

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 07-200187

(43)Date of publication of application : 04.08.1995

(51)Int.Cl.

G06F 3/06
G06F 13/10

(21)Application number : 05-351130

(71)Applicant : HITACHI LTD

(22)Date of filing : 30.12.1993

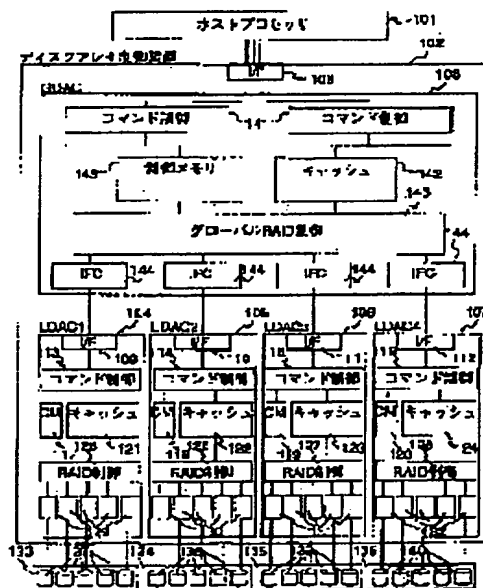
(72)Inventor : TAKAMOTO YOSHIFUMI
TSUNODA HITOSHI

(54) DISK ARRAY DEVICE

(57)Abstract:

PURPOSE: To provide a disk array device which can carry out fast input/output operation, facilitates disk management, and causes no decrease in reliability even when many disk drives are connected.

CONSTITUTION: This disk array device is equipped with a global disk array controller 103 which puts local disk array controllers 104 to 107, where plural disk drives 137 to 140 are connected, further in a disk array. Each disk array controller performs RAID control over plural disk drives in its machine. The global disk array controller performs RAID control over the local disk array controllers. Consequently, highly reliable disk devices are obtained by fastness by the division of data and global parity which are generated by the local disk arrays and global disk array and stored in the local disk arrays.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C): 1998,2003 Japan Patent Office

THIS PAGE BLANK (USPTO)

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平7-200187

(43) 公開日 平成7年(1995)8月4日

(51) Int.Cl. ⁶	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 3/06	5 4 0			
13/10	3 4 0 A	8327-5B		

審査請求 未請求 請求項の数13 F D (全 28 頁)

(21) 出願番号 特願平5-351130

(22) 出願日 平成5年(1993)12月30日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 高本 良史

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 角田 仁

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

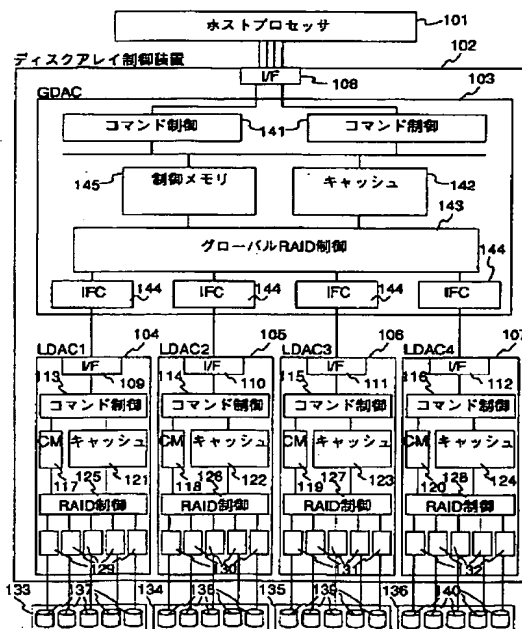
(74) 代理人 弁理士 矢島 保夫

(54) 【発明の名称】 ディスクアレイ装置

(57) 【要約】 (修正有)

【目的】多数のディスクドライブを接続しても、高速な入出力が可能であり、またディスク管理も容易で信頼性を低下させることもないディスクアレイ装置を提供する。

【構成】複数のディスクドライブ137-140が接続された複数のローカルディスクアレイ制御装置104-107を、さらにディスクアレイ化するためのグローバルディスクアレイ制御装置103を備えている。ローカルディスクアレイ制御装置は、自機内の複数のディスクドライブをRAID制御する。グローバルディスクアレイ制御装置は、ローカルディスクアレイ制御装置に対しRAID制御する。この結果、データの分割による高速性と、ローカルディスクアレイとグローバルディスクアレイが生成しローカルディスクアレイに格納するグローバルバリティとにより高信頼なディスク装置が得られる。



(2)

特開平7-200187

1

【特許請求の範囲】

【請求項1】ホストプロセッサから転送されたデータを所定数のグループに分割し各グループのデータを並列に出力する手段と、

並列に出力された各グループのデータを、グループ単位でそれぞれ入力し格納する複数のローカルなディスクアレイ装置とを備えたことを特徴とするディスクアレイ装置。

【請求項2】複数のローカルなディスクアレイ装置と、ホストプロセッサから転送されたデータを、前記複数のローカルなディスクアレイ装置のうちの何れに出力するかを選択し、選択したディスクアレイ装置に該データを出力する手段とを備えたことを特徴とするディスクアレイ装置。

【請求項3】ホストプロセッサから転送されたデータを所定数のグループに分割し各グループのデータを並列に出力する手段と、

前記複数のグループ間のパリティを生成して、前記各グループのデータと並列に出力する手段と、

並列に出力された各グループのデータを、グループ単位でそれぞれ入力し格納する複数のローカルなディスクアレイ装置と、

前記パリティを入力し格納するローカルなディスクアレイ装置とを備えたことを特徴とするディスクアレイ装置。

【請求項4】複数のディスクドライブを含み、ホストプロセッサから転送されたデータと前記データから生成されたパリティを、前記ディスクドライブに格納するディスクアレイ装置であって、

前記複数のディスクドライブが、それぞれ複数のディスクドライブを含む複数の論理グループに分けられているとともに、

前記ホストプロセッサから転送されたボリウム内アドレスとデータ長に基づいて、前記論理グループの一つまたは複数を選択し、選択した論理グループに、格納すべきデータを転送する手段と、

各論理グループごとに設けられ、前記転送されたデータから論理グループ内のパリティを生成し、自論理グループ内の複数のディスクドライブに前記データと生成したパリティとを格納する手段とを備えたことを特徴とするディスクアレイ装置。

【請求項5】複数のディスクドライブを含み、ホストプロセッサから転送されたデータと前記データから生成されたパリティを、前記ディスクドライブに格納するディスクアレイ装置であって、

前記複数のディスクドライブが、それぞれ複数のディスクドライブを含む複数の論理グループに分けられているとともに、

前記ホストプロセッサから転送されたボリウム内アドレスとデータ長に基づいて、前記論理グループの一つまた

2

は複数を選択し、選択した論理グループに、格納すべきデータを転送する手段と、

前記選択した各論理グループに転送するデータを用いて論理グループ間パリティを生成し、前記選択した論理グループとは異なる論理グループを選択し、選択した論理グループに、格納すべきデータとして前記論理グループ間パリティを転送する手段と、

各論理グループごとに設けられ、前記転送された格納すべきデータから論理グループ内のパリティを生成し、自論理グループ内の複数のディスクドライブに前記データと生成したパリティとを格納する手段とを備えたことを特徴とするディスクアレイ装置。

【請求項6】請求項4または5に記載のディスクアレイ装置であって、

前記論理グループ内のパリティが格納されたディスクドライブを除く他のディスクドライブに対し、各ディスクドライブへの格納単位ごとに順次連続した前記ボリウム内アドレスを割り付け、さらに前記ボリウム内アドレスは複数の前記論理グループに渡って順次水平に割り付けることを特徴とするディスクアレイ装置。

【請求項7】請求項6に記載のディスクアレイ装置であって、

前記論理グループにおける前記ボリウム内アドレスの連続部分を前記論理グループの最小格納単位とし、ホストプロセッサから転送されたデータを前記最小格納単位ごとに複数の前記論理グループに水平に分割し各論理グループに転送することを特徴とするディスクアレイ装置。

【請求項8】請求項7に記載のディスクアレイ装置であって、

前記各論理グループ内では、前記最小格納単位で転送されたデータを、ディスクドライブの格納単位ごとに水平に複数のディスクドライブに格納することを特徴とするディスクアレイ装置。

【請求項9】請求項8に記載のディスクアレイ装置であって、

前記論理グループの前記最小格納単位を満たしたデータから、順次、前記データに対応する前記論理グループに転送することを特徴とするディスクアレイ装置。

【請求項10】請求項6に記載のディスクアレイ装置であって、

前記論理グループにおける前記ボリウム内アドレスの連続部分のデータが前記ホストプロセッサから転送されたとき、前記転送されたデータからパリティを生成し、前記データと前記パリティとを並列に前記論理グループ内のディスクドライブに格納することを特徴とするディスクアレイ装置。

【請求項11】請求項4または5に記載のディスクアレイ装置であって、

前記論理グループ内の何れかのディスクドライブに障害が発生したとき、前記論理グループ内に格納されたパリ

(3)

特開平7-200187

3

ティと障害ディスクドライブ以外のディスクドライブのデータとから、前記障害ディスクドライブのデータを生成することを特徴とするディスクアレイ装置。

【請求項12】請求項5に記載のディスクアレイ装置であって、前記論理グループの何れかに障害が発生したとき、前記論理グループ間バリティと障害を起こした論理グループ以外の論理グループのデータとから、前記障害を起こした論理グループのデータを生成することを特徴とするディスクアレイ装置。

【請求項13】複数のディスクドライブを含み、ホストプロセッサから転送されたデータと前記データから生成されたバリティを、前記ディスクドライブに格納するディスクアレイ装置であって、

前記複数のディスクドライブが、それぞれ複数のディスクドライブを含む複数の論理グループに分けられているとともに、各論理グループごとに論理グループ制御装置が設けられており、

該論理グループ制御装置は、ホストプロセッサから転送された入出力コマンドが、自装置に対応する論理グループに対する入出力要求かどうかを判定し、自装置に対応する論理グループに対する入出力要求である場合はその入出力コマンドを実行し、そうでなければその入出力コマンドをキャンセルする手段を備えたことを特徴とするディスクアレイ装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、コンピュータシステムなどに用いるディスクファイルシステムに関し、特に高速でかつ高信頼なディスクアレイ装置に関する。

【0002】

【従来の技術】一般的に、コンピュータシステムは、プロセッサと2次記憶装置とを備えている。多く使用される2次記憶装置として、磁気ディスク装置がある。磁気ディスク装置において、現在その容量の伸び率は極めて高いが、メカニカル動作を伴う磁気ディスク装置の性能はプロセッサ性能の伸び率ほど高くない。その課題を解決する方式として、ディスクアレイが提案された。

【0003】ディスクアレイに関する代表的な論文として、D. Patterson, G. Gibson, and R. H. Kartzらによる「A Case for Redundant Arrays of Inexpensive Disks (RAID), in ACM SIGMOD Conference, Chicago, IL, (June 1988)」がある。RAIDとは、複数のディスクドライブにデータを分散して配置することでアクセス時間を短縮し、かつ、誤り訂正用のバリティあるいはECCと呼ばれる冗長データを格納することで信頼性も高めることができる技術である。つまり、複数のディスクドライブに対して並列に入出力を行うことができることによる高速性と、何れかのディスクドライブに障害が発生したときでもバリティと障害ディスクドライブ以外のデータから障害デ

4

ィスクドライブのデータを回復することができる技術である。

【0004】この技術を実現するための特許出願もなされており、例えば、ジャイシヤンカー ムーセダス ミノン、ジェイムス マシューズ カーソン (インターナショナル ビジネス マシーンズ CORP) らによる特開平4-230512号がある。この公報には、ディスクアレイにおける課題の一つであるバリティの更新を高速にする技術が開示されている。

10 【0005】ディスクアレイでは、データ書き込み時に、データ自体とバリティの更新を行わなくてはならない。バリティの更新のためには、更新前のデータとバリティが必要となるケースがあり、そのためそれらのデータやバリティを事前に読みだすオーバーヘッドと元のデータやバリティの位置に新しいデータやバリティを書き込むオーバーヘッドがかかる。これをライト・ペナルティと呼ぶ。上記公報には、新しいデータやバリティを元の位置に記録せず、ディスクドライブの空き領域に格納することで、ライト・ペナルティによるオーバーヘッドを少なくする技術が開示されている。

20 【0006】他に、ディスクアレイの実装を容易にするための技術として、特開平5-46524号に開示されたものがある。この公報に記載のディスクアレイ装置は、ディスクアレイの実装を容易にするために、複数のディスクを共通のバスに接続し、そのバスに接続された上位コントローラがディスクアレイ制御を行うものである。

【0007】

30 【発明が解決しようとする課題】ディスクアレイコントローラの基本的な機能として、ホストプロセッサからの入出力要求を解釈/実行、データの分割、バリティの生成、キャッシュメモリ管理、およびディスクドライブの起動/終了処理がある。このように、ディスクアレイコントローラは、通常のディスクコントローラに比べて多くの処理を行わなければならない。

【0008】例えば、2台のドライブ (データを格納する1台のディスクドライブとバリティを格納する1台のディスクドライブ) でRAID5のディスクアレイ装置を構成したとする。このとき、新データの書き込み時には、旧データと旧バリティをそれぞれ2台のディスクドライブから読みだし、その後、新データと新バリティを前記2台のディスクドライブに出力しなければならない。つまり、1つの出力処理で4回のディスク入出力が発生することになり、これがライト・ペナルティになる。その他にも、ディスクアレイコントローラ特有の処理として、データの分割、バリティの生成などを行なう必要がある。

【0009】従来のディスクアレイ装置は、一つのディスクアレイコントローラで複数のディスクドライブを制御していた。そのような構成で多数のディスクドライブ

(4)

特開平7-200187

5

を制御するためには、高速なディスクアレイコントローラを設けるか、同様のディスクアレイコントローラを複数設ける必要がある。

【0010】しかし、高速なディスクアレイコントローラは非常に高価であり、これはディスクアレイ装置全体のコストを上げることにつながる。また、同様のディスクアレイコントローラを複数設けると、利用者から複数のディスク装置に見えてしまい、ディスク管理を複雑にする欠点がある。

【0011】本発明の目的は、ディスクアレイ装置の改良にある。また、本発明の目的は、多数のディスクドライブを接続しても、高速な入出力が可能であり、コストを上げることがなく、またディスク管理も容易で信頼性を低下させることもないディスクアレイ装置を提供することにある。

【0012】

【課題を解決するための手段】本発明では、複数のディスクドライブが、それぞれ複数のディスクドライブを含む複数の論理グループに分けられており、各論理グループは、例えば各論理グループごとに設けられたディスクアレイ制御装置によりそれぞれ管理される。そして、ホストプロセッサから転送されたボリューム内アドレスとデータ長に基づいて、前記論理グループの一つまたは複数を選択し、前記選択された論理グループごとに格納されるデータからパリティを生成し、前記論理グループ内のディスクドライブに前記データとパリティを格納するようにする。

【0013】さらに、前記論理グループ内のパリティが格納されたドライブを除く他のディスクドライブに対し、ディスクドライブの格納単位ごとに順次連続した前記ボリューム内アドレスを割り付け、さらに前記ボリューム内アドレスは複数の前記論理グループに渡って順次水平に割り付ける。

【0014】

【作用】複数のディスクドライブから構成される複数の論理グループに分け、前記論理グループごとに例えばディスクアレイ制御装置を設けて管理するので、小数のディスクドライブを管理する安価なディスクアレイコントローラになり、また、前記論理グループ内のパリティが格納されたドライブを除く他のディスクドライブに対し、ディスクドライブの格納単位毎に順次連続した前記ボリューム内アドレスを割り付け、さらに前記ボリューム内アドレスは複数の前記論理グループに渡って順次水平に割り付けることで、利用者からは単一のディスクドライブとして管理することができるようになる。

【0015】

【実施例】以下、図面を用いて、本発明の実施例を詳細に説明する。

【0016】【実施例1】図1は、本発明の第1の実施例に係るディスクアレイ装置の全体構成図を示したもの

6

である。

【0017】ディスクアレイ制御装置(102)は、ホストプロセッサ(101)と接続されている。ディスクアレイ制御装置(102)は、複数のディスク装置(133, 134, 135, 136)に接続されている。各ディスク装置(133, 134, 135, 136)は、それぞれ複数のディスクドライブ(137, 138, 139, 140)を備えている。

【0018】ディスクアレイ制御装置(102)は、インタフェース制御部(108)と、グローバルディスクアレイ制御装置(103)と、複数のローカルディスクアレイ制御装置(104, 105, 106, 107)とから構成されている。

【0019】インタフェース制御部(108)は、ホスト(101)との間のデータあるいはコマンドの転送プロトコル制御を行う。グローバルディスクアレイ制御装置(103)は、複数のローカルディスクアレイ制御装置(104, 105, 106, 107)に対して、データの分割、パリティの生成/格納、あるいはそれらの複数のローカルディスクアレイ制御装置(104, 105, 106, 107)から転送されたデータのマージ処理などを行う。

【0020】1つのローカルディスクアレイ制御装置(104, 105, 106, 107)は、1つの論理グループに相当する。各ローカルディスクアレイ制御装置(104, 105, 106, 107)は、グローバルディスクアレイ制御装置(103)から受けたデータの分割格納やパリティの生成/格納などを行う。論理グループとは、RAID制御の単位である。すなわち、論理グループを構成するローカルディスクアレイ制御装置は、上位から受けたデータを分割しパリティを生成して複数のドライブに格納するが、それら複数のドライブを含むRAID制御の単位を論理グループという。

【0021】グローバルディスクアレイ制御装置(103)と複数のローカルディスクアレイ制御装置(104, 105, 106, 107)は共に、ディスク装置(133, 134, 135, 136)の障害やディスク装置(133, 134, 135, 136)内のディスクドライブ(137, 138, 139, 140)の障害時に、パリティを使用して障害回復が可能である。

【0022】本実施例の特徴の一つは、ローカルディスクアレイ制御装置(104, 105, 106, 107)を制御するグローバルディスクアレイ制御装置(103)を設けていることである。これらの関連した詳細動作は後で説明する。

【0023】グローバルディスクアレイ制御装置(103)は、コマンド制御部(141)、制御メモリ(145)、キャッシュメモリ(142)、グローバルRAID制御部(143)、およびインタフェースコントローラ(144)から構成されている。

(5)

特開平 7-200187

7

【0024】コマンド制御部(141)は、ホストプロセッサ(101)からの入出力要求を解釈し、制御メモリ(145)内のキューに入出力要求をキューイングする。制御メモリ(145)内には、ディスクアレイ装置全体を管理するための制御情報が格納されている。キャッシュメモリ(142)内には、ディスクアレイ装置(102)が入出力を行ったデータを、キャッシュメモリ(142)の容量の範囲内で格納しておく。頻繁にアクセスされるデータは、キャッシュメモリ(142)に残っている確率が高くなる。

【0025】グローバルRAID制御部(143)は、複数のローカルディスクアレイ制御装置(104, 105, 106, 107)を、ホスト101からは単体ディスクまたは複数ディスクとして取り扱うことができるような制御を行う。インタフェースコントローラ(144)は、ローカルディスクアレイ制御装置(104, 105, 106, 107)との間のデータあるいはコマンドの転送制御を行う。

【0026】以上が本実施例のディスクアレイ装置の構成概要であり、詳細な説明を以下に述べる。

【0027】図2は、本発明における動作の概要の一例を示しており、本発明の特徴と効果の一例を以下に述べる。

【0028】本動作概要は、ホストプロセッサ(101)からデータ(223)がディスク装置(133, 134, 135)内のディスクドライブに格納されるまでの動作とデータの流れを示している。本例は、ローカルディスクアレイ制御装置(104)がRAID3の動作を行なったケースである。

【0029】ホストプロセッサ(101)から転送されたデータ(223)は、ディスクアレイ制御装置(102)内のグローバルディスクアレイ制御装置(103)で、ローカルディスクアレイ制御装置(104, 105, 106)の台数分(3つ)に分割される(201)。これらのデータ(219, 220, 221)は、並列にローカルディスクアレイ制御装置(104, 105, 106)に転送される。

【0030】ローカルディスクアレイ制御装置(104, 105, 106)にデータが転送されると、ローカルディスクアレイ制御装置(104, 105, 106)内で、各ローカルディスクアレイ制御装置(104, 105, 106)に接続されているディスクドライブ数からパリティ格納用ディスクドライブ数を除いた台数分にデータの分割が行われ(207, 209, 211)、さらに同時にグローバルディスクアレイ制御装置(103)から転送されたデータのパリティが生成される(208, 210, 212)。ローカルディスクアレイ制御装置(104, 105, 106)内で分割されたデータとパリティは並列にディスクドライブに格納される。

【0031】ローカルディスクアレイ制御装置(10

8

4, 105, 106)は、それぞれ単独で動作可能な機構を持っているため、全てのディスクドライブの並列動作が可能となる。こういった構成により、大規模なデータの高速な入出力が可能となる。

【0032】図3は、本発明における動作の概要の一例を示しており、本発明の特徴と効果の一例を以下に述べる。

【0033】本動作概要は、ホストプロセッサ(101)からデータ(223)がディスク装置(133)内のディスクドライブに格納されるまでの動作とデータの流れを示している。本例は、ローカルディスクアレイ制御装置(104)がRAID5の動作を行なったケースである。

【0034】ホストプロセッサ(101)からデータ(223)が転送されると、ディスクアレイ制御装置(102)内のグローバルディスクアレイ制御装置(103)では、そのデータを格納すべきローカルディスクアレイ制御装置(104)の選択が行なわれる。

【0035】ローカルディスクアレイ制御装置(104)にデータが転送されると、ローカルディスクアレイ制御装置(104)内で、ローカルディスクアレイ制御装置(104)に接続されているディスクドライブ数からパリティ格納用ディスクドライブ数を除いた台数分にデータの分割が行われ(207)、さらに同時にグローバルディスクアレイ制御装置(103)から転送されたデータのパリティの一括生成が行なわれる(208)。

【0036】通常RAID5は、ブロック単位の入出力を行なうため、複数ブロックの格納時には各ブロック毎にライト・ペナルティが生じる。しかし、本例に示すように、連続して複数のブロックの書き込み時には、本発明のように、連続してアドレスを設定することで、一括してパリティの生成を行なうことができるようになる。

【0037】ローカルディスクアレイ制御装置(104)内で分割されたデータとパリティは並列にディスクドライブに格納される。こういった構成により、短いデータのスループットを向上することができる。

【0038】図4は、本発明における動作の概要の一例を示しており、本発明の特徴と効果の一例を以下に述べる。

【0039】本動作概要は、ホストプロセッサ(101)からデータ(223)がディスク装置(133, 134, 135, 136)内のディスクドライブ(137, 138, 139, 140)に格納されるまでの動作とデータの流れを示している。

【0040】ホストプロセッサ(101)から転送されたデータ(223)は、ディスクアレイ制御装置(102)内のグローバルディスクアレイ制御装置(103)で、ローカルディスクアレイ制御装置(104, 105, 106, 107)の台数からグローバルパリティ格

(6)

特開平 7-200187

9

納用のローカルディスクアレイ制御装置(107)の台数を除いた数に分割される(201)。同時に、ホストプロセッサ(101)から転送されたデータ全体のグローバルパリティが生成される(202)。これらのデータは、並列にローカルディスクアレイ制御装置(104, 105, 106, 107)に転送される。

【0041】一般に、ディスクアレイは、信頼性の向上のために、データとは異なるパリティを生成しディスクドライブに格納しておく。もし、ディスクドライブに障害が発生しても、パリティと障害が発生したディスクドライブ以外のディスクドライブのデータとから、障害が発生したディスクドライブのデータを、演算により求めることができる。

【0042】ローカルディスクアレイ制御装置(104, 105, 106, 107)にデータが転送されると、各ローカルディスクアレイ制御装置(104, 105, 106, 107)内で、ローカルディスクアレイ制御装置(104, 105, 106, 107)に接続されているディスクドライブ(137, 138, 139, 140)の数からパリティ格納用ディスクドライブを除いた台数分にデータの分割が行われる(207, 209, 211, 213)。さらに同時に、グローバルディスクアレイ制御装置(103)から転送されたデータのパリティが生成される(208, 210, 212, 214)。ローカルディスクアレイ制御装置(104, 105, 106, 107)内で分割されたデータとパリティは、並列にディスクドライブ(137, 138, 139, 140)に格納される。

【0043】ローカルディスクアレイ制御装置(104, 105, 106, 107)は、それぞれ単独で動作可能な機構を持っているため、全てのディスクドライブ(137, 138, 139, 140)の並列動作が可能となる。こういった構成により、大規模なデータの高速な入出力が可能となる。

【0044】さらには、通常のディスクアレイのパリティ以外に、ローカルディスクアレイ制御装置(104, 105, 106)間のパリティ(グローバルパリティ)を生成／格納していることから高い信頼性を得ることができる。つまり、各ローカルディスクアレイ制御装置(104, 105, 106, 107)での回復処理と、上位のグローバルディスクアレイ制御装置(103)での回復処理といったように、本発明では障害回復処理が多層化されることによる。

【0045】図5は、図2から図4で述べた特徴と効果とは別の特徴と効果の一例を示している。具体的には、本発明の入出力要求と終了報告の流れを示している。

【0046】ホストプロセッサ(101)からデータが転送されると、グローバルディスクアレイ制御装置(103)はそのデータを分割しグローバルパリティを生成して、下位のローカルディスクアレイ制御装置(104, 105, 106, 107)の領域の単一制御やユーザの要求

10

4, 105, 106, 107)に並列に入出力を行なう(301)。ローカルディスクアレイ制御装置(104, 105, 106, 107)は、グローバルディスクアレイ制御装置(103)から転送されたデータを分割しパリティを生成して、ディスクドライブ(137, 138, 139, 140)に並列に入出力を行なう(306, 307, 308, 309)。

【0047】いま仮にローカルディスクアレイ制御装置(104, 105, 106, 107)の数を m とし、各ローカルディスクアレイ制御装置(104, 105, 106, 107)に接続されたドライブ数を d とすると、全部で md 台のディスクドライブがあることになる。もし、一つのディスクアレイ制御装置にこの md 台の全てのディスクドライブが接続されたとすると、 md 回の入出力を一つの制御装置がまかなうことになり、非常に高速な制御装置が必要となる。

【0048】しかし、本発明では、各ローカルディスクアレイ制御装置(104, 105, 106, 107)毎に d 回と少ない入出力回数を処理するだけでよく、ローカルディスクアレイ制御装置(104, 105, 106, 107)に高速な処理は要求されない。

【0049】図6は、コマンド制御部(141)の構成を示している。コマンド制御部(141)は、インタフェース制御部(108)とのコマンドやデータの転送制御を行うアダプタ(401)、プロセッサ(402)、DMAC(403)、メモリ(404)、バス(407)、およびバッファ(408)から構成されている。

【0050】メモリ(404)内には、コマンド制御を行うマイクロプログラム(405, 406)が格納されている。具体的には、コマンド受付処理プログラム(405)と終了処理プログラム(406)である。これらの詳細は後で述べる。

【0051】マイクロプログラム(405, 406)は、共通バス(408)を介して、プロセッサ(402)で実行される。コマンド制御部(141)と外部の制御部とのデータ転送は、ダイレクトメモリアクセスコントローラであるDMAC(403)が、プロセッサ(402)とは独立に、実行する。コマンド制御部(141)から制御メモリ(145)やキャッシュにデータやコマンドを転送する際は、バッファ(408)とバス(409)を介して行われる。

【0052】図7は、制御メモリ(145)に格納されているテーブルやリストを示している。502は仮想ドライブ管理テーブル、503はI/O管理リスト、504はローカルディスクアレイ管理テーブル、505はキャッシュ管理リスト、506はコマンドキューリストである。

【0053】仮想ドライブ管理テーブル(502)は、複数のローカルディスクアレイ制御装置(104, 105, 106, 107)の領域の単一制御やユーザの要求

(7)

特開平7-200187

11

に合わせて設定する領域を管理している。I/O管理リスト(503)は、ホストプロセッサ(101)から受け取った入出力要求を入出力が環境するまで入出力状態等と共に管理するテーブルである。

【0054】ローカルディスクアレイ管理テーブル(504)は、グローバルディスクアレイ制御装置(103)に接続されているローカルディスクアレイ制御装置(104, 105, 106, 107)を管理するテーブルであり、RAIDレベル、容量、およびパーティション構成等が格納されている。キャッシュ管理リスト(505)は、キャッシュメモリ(142)内に格納されているデータとそのアドレスが格納されている。コマンドキューリスト(506)は、グローバルディスクアレイ制御装置(103)がコマンドに対してタグを付加する為の管理情報が格納されている。

【0055】制御メモリ145のポインタ501からポイントされているこれらのテーブルやリストについての詳細な構成は後で述べる。

【0056】図8は、ローカルディスクアレイ管理テーブル(504)の構成を示したものである。

【0057】カラム601には、ローカルディスクアレイの識別子が格納されている。カラム602には、ローカルディスクアレイのパーティション識別子が格納されている。このカラムにより、単一のローカルディスクアレイ領域を複数に区切ることができる。これは、ディスクアレイが複数のディスクドライブから構成されていることから、単一領域として使用するには大きすぎることがある為である。このカラムはユーザが任意に指定可能である。

【0058】カラム603にはバス識別子が格納されている。グローバルディスクアレイ制御装置(103)がローカルディスクアレイ制御装置(104, 105, 106, 107)に対して入出力要求を発行するときに、このバス識別子を参照することで、ローカルディスクアレイ制御装置(104, 105, 106, 107)の位置を知ることができる。カラム604は、各ローカルディスクアレイ制御装置(104, 105, 106, 107)のドライブ保有数を示している。グローバルディスクアレイ制御装置(103)がグローバルパリティを作成する場合のパリティ位置を算出するときなどに使用する。

【0059】カラム605は、各ローカルディスクアレイ制御装置(104, 105, 106, 107)の制御しているRAIDレベルを示している。これもカラム604と同じように、グローバルパリティを作成する場合のパリティ位置を算出するときなどに使用する。カラム606には、各ローカルディスクアレイ制御装置(104, 105, 106, 107)のパリティを格納するディスク数が格納されている。これもカラム605と同じように、グローバルパリティを作成する場合のパリティ

12

位置を算出するときなどに使用する。

【0060】カラム607には、各ローカルディスクアレイ制御装置(104, 105, 106, 107)のパーティション毎のストライピングサイズが格納される。ストライピングサイズは、ローカルディスクアレイ制御装置(104, 105, 106, 107)の最小入出力単位を意味する。例えば、ストライピングサイズが4(KB)では、4KB以内のデータ長のデータは、単一ディスクドライブに格納される。4KBを超える長さのデータは、4KB毎に次のディスクドライブに分割して格納される。

【0061】カラム608には、パーティション毎の容量が格納される。カラム609には、ステータスが格納されており、例えば障害の発生等の各ローカルディスクアレイ制御装置(104, 105, 106, 107)の状態が格納される。

【0062】図9は、仮想ドライブ管理テーブル(502)の構成を示している。仮想ドライブ管理テーブルは、ホストプロセッサ(101)から認識するディスクドライブと物理的なディスクドライブとのマッピングを行うために設けられる。

【0063】カラム701には、仮想ドライブの識別子が格納される。カラム702から705までは、仮想ドライブ識別子で示される仮想ドライブが、ローカルディスクアレイ制御装置(104, 105, 106, 107)のどのパーティションから構成されるかを示している。例えば、仮想ドライブVOL1は、ローカルディスクアレイ制御装置(104, 105, 106, 107)のパーティション1から構成される。

【0064】カラム706には、仮想ドライブ容量が格納される。カラム707には、グローバルRAIDレベルが格納されている。このグローバルRAIDレベルは、ローカルディスクアレイ制御装置(104, 105, 106, 107)のRAIDレベルとは無関係に設定可能である。カラム708には、パリティを格納するディスク数が格納されている。カラム709には、グローバルディスクアレイ制御装置(103)のストライピングサイズが格納される。カラム710には、ステータスが格納されており、例えば障害の発生等の仮想ドライブ管理情報が格納される。

【0065】図10は、図7のような仮想ドライブ管理テーブル(502)を用いることにより、ホストプロセッサ(101)からどのように仮想ドライブが見えるか、および物理的にどのようにデータが配置されているかを示す概念図である。

【0066】ホストプロセッサ(101)から入出力要求(801, 802, 803)が発行される。このとき、ホストプロセッサ(101)は、仮想ドライブ(804, 805, 806)に対して入出力を行なっている。このような仮想ドライブに対する入出力に対し、物

(8)

特開平7-200187

13

理的には、ディスク装置(133, 134, 135, 136)にデータが分散配置される。ホストプロセッサ(101)から見える仮想ドライブ(804, 805, 806)は、物理的には、LDRV1(807), LDRV2(808), LDRV3(809)のように、ディスク装置(133, 134, 135, 136)を横切る形態で分散されている。

【0067】仮想ドライブ管理テーブル(502)の内容を変更することで、必ずしもディスク装置(133, 134, 135, 136)の全てに渡って分散配置される構造にはならない。一つのローカルディスクアレイ領域を一つの仮想ドライブと定義することも可能である。さらに、810, 811, 812は、ローカルディスクアレイ制御装置(104, 105, 106, 107)が生成/格納したパリティであり、813はグローバルディスクアレイ制御装置(103)が生成/格納したパリティを示している。こういったテーブルにより、ユーザの要求する容量を満足し、また、RAIDレベル等を自由に設定することができるため、性能における柔軟性もある。

【0068】図11は、I/O管理リスト(503)の構成を示している。

【0069】制御メモリ(145)からポイントされているアドレスは、さらに二つのポイントを示している。一つは次の入出力要求のリストが格納されている領域へのポイント(901)であり、もう一つは入出力要求が格納されている領域を示すポイント(902)である。909は、リストの最後であることを示す識別子である。

【0070】格納されている入出力要求は、903から908の範囲で示されている。903は何処のホストプロセッサからの要求であるかを識別するためのホスト識別子、904は入出力を受け付けた時間、905は入出力の状態が格納されている。入出力の状態とは、例えば、入出力実行中などである。906は、ホストプロセッサから転送されたコマンドが格納されている。

【0071】907は、ホストからコマンドが転送される際に付加されているタグが格納されている。このタグは、入出力を発行したホストプロセッサが、どの入出力要求が完了したのかを識別するためのものである。これは、ホストプロセッサが同一デバイスに対して複数の入出力要求を発行できる場合に付加される。もしこのタグがなければ、ホストプロセッサは、同一デバイスから終了通知が帰ってきても、どの入出力に対する完了かを識別することができないからである。

【0072】908は新タグが格納されている。この新タグ(908)は、グローバルディスクアレイ制御装置(103)が設定するもので、ホストからのタグ(907)はホストが入出力を識別するものであるが、新タグ(908)はグローバルディスクアレイ制御装置(10

14

3)が同一ローカルディスクアレイ制御装置(104, 105, 106, 107)に対して複数の入出力要求を発行したときに使用するものである。

【0073】この新タグ(908)が必要な理由は、ホストが識別する仮想ドライブと、グローバルディスクアレイ制御装置(103)が識別するローカルディスクアレイ制御装置(104, 105, 106, 107)とが同一でないことによる。つまり、タグを付加する側はドライブ毎にユニークなタグを付加する。しかし、このタグをそれを受け取った装置がホストの認識するドライブとは異なるドライブに使用すると、ドライブ毎のタグのユニークさが失われることになるためである。そのために、グローバルディスクアレイ制御装置(103)が使用するタグとしては、ホストから転送されたタグは使用せず、グローバルディスクアレイ制御装置(103)が管理できる新タグ(908)を使用する必要がある。

【0074】図12は、キャッシュ管理リスト(505)の構造を示している。キャッシュ管理リスト(505)は、キャッシュメモリ(142)内にキャッシングされているデータの場所やディスクドライブ(137, 138, 139, 140)との対応等を管理するためのリストである。

【0075】制御メモリ(145)からポイントされているアドレスは、さらに二つのポイントを示している。一つはキャッシュメモリのフリーポイント(1001)であり、もう一つは使用中ポイント(1002)である。キャッシュメモリ(142)は、ブロック単位で管理されており、前記フリーポイント(1001)と使用中ポイント(1002)が指すリストの一つ一つ(1002-1006, 1007-1011)は固定のブロックに対応している。

【0076】フリーポイント(1001)からポイントされているリスト(1002-1006)は、キャッシュメモリ(142)の未使用領域(ブロック)を管理するために使用される。もし、新規にキャッシュメモリ(142)の領域を取得するときには、このフリーポイント(1001)からポイントされているリスト(1002-1006)の中から必要なだけのリストを取得した後、後述の使用中ポイント(1002)がポイントしているリスト内に入れる操作を行う。

【0077】各リストには、次のリストをポイントするエリアと、該リストがディスク上のどここのデータ(BLK#)を保持しているかを示すエリアと、キャッシュメモリ(142)の何処に格納されているかを示すポイントとが、格納されている(1043-1045)。フリーポイントがポイントするリストは未使用リストがチェインされていることから、キャッシュメモリ(142)へのポイントの値は意味を持たない。

【0078】1007はハッシュリストである。ハッシュリスト(1007)は、要求されたデータがキャッシ

(9)

特開平7-200187

15

ュメモリ(142)内に格納されているかどうかを高速に検索する目的で設けられている。要求するディスクドライブのブロック番号BLK#をキーとしてハッシングすれば、簡単に、検索しなければならない候補を絞ることが可能となる。

【0079】図13は、コマンドキューリスト(506)の構成を示している。

【0080】1101は、各ローカルディスクアレイ制御装置(104, 105, 106, 107)に対応したコマンドキューへのポイントが格納されている。1102と1103は、ローカルディスクアレイ制御装置(104)に対応するフリー新タグ(1110-1114)と使用新タグ(1115-1118)のリストをポイントしている。1104と1105は、ローカルディスクアレイ制御装置(105)に対応するフリー新タグ(1119-1122)と使用新タグ(1123-1127)のリストをポイントしている。1106と1107は、ローカルディスクアレイ制御装置(106)に対応するフリー新タグ(1128-1132)と使用新タグ(1133-1136)のリストをポイントしている。1108と1109は、ローカルディスクアレイ制御装置(107)に対応するフリー新タグ(1137-1141)と使用新タグ(1142-1145)のリストをポイントしている。

【0081】各リストには、次のリストへのポイントと、新タグ値と、ホストプロセッサ(101)から転送された入出力コマンド(ホストからのタグも含む)とが、格納されている。フリー新タグ(1110-1114, 1119-1122, 1128-1132, 1137-1141)中のコマンドエリア(cmd)の値は意味を持たない。使用新タグのリストには、それぞれホストプロセッサ(101)からの入出力要求の一つが格納されている。使用新タグ(1115-1118, 1123-1127, 1133-1136, 1142-1145)は、コマンドの実行終了時に、使用新タグリストからフリー新タグリストへ移される。各ローカルディスクアレイ制御装置(104, 105, 106, 107)に対応するコマンドキューリストは、それぞれ重複しないタグ値を持っている。このコマンドキューリスト(506)により、前述のホストプロセッサ(101)から転送されたタグを新タグに変換することが可能となる。

【0082】図14は、コマンド受付処理プログラム(405)のフローチャートを示している。コマンド受付処理プログラム(405)は、コマンド制御部(145)内のメモリ(404)に格納されているマイクロプログラムである。

【0083】ステップ1201ではキャッシュ確保可能かどうかをチェックする。もし可能であればステップ1202へ移り、そうでなければステップ1212へ移りコマンド異常終了処理を行う。通常、キャッシュメモリ

16

(142)に領域が確保できないことはない。コマンド異常終了処理が実行されるのは、キャッシュメモリ(145)がアクセス不可能な状態、例えば障害が発生したときである。

【0084】ステップ1202では、キャッシュメモリ(145)の所要分領域を確保する。ステップ1203では、キャッシュ管理リスト(505)のフリーポイント(1001)からポイントされているリストから、所要分のリストを使用中ポイント(1002)がポイントしているリスト内に入れる。ステップ1204では、I/O管理リスト(503)中に空きリストが存在するかどうか調べる。もし、空きリストが存在すればステップ1205に進み、そうでなければステップ1212に進み異常終了処理を行う。この状態は、ディスクアレイ制御装置(102)内にホストプロセッサ(101)からのコマンドを受け付けるためのキューが無くなったことを意味する。

【0085】ステップ1205では、I/O管理リスト(503)に新しいリストを追加する。これにより、当該コマンドがディスクアレイ制御装置(102)の処理対象となる。ステップ1206では、ホストIDをI/O管理リスト(503)のエリア(903)に格納する。ステップ1208では、コマンド受付時間をI/O管理リスト(503)のエリア(904)に格納する。この時間は、もしディスクアレイ制御装置(102)に障害が発生した場合などの保守に使用することができる。

【0086】ステップ1209では、コマンドをI/O管理リスト(503)のエリア(906)に格納する。このとき、I/O管理リスト(503)のエリア907にもホストプロセッサ(101)から転送されたタグを格納する。これは、入出力要求完了時に、ホストプロセッサ(101)に対して完了情報と必要であればデータなどと共にホストプロセッサ(101)へ転送する際に参照される。ステップ1210では、ステップ1202で取得したキャッシュメモリ(142)の領域にコマンドを転送する。ここでコマンドとは、READ/WRITEなどのディスク装置に対する指示と、WRITE要求であればデータも同様に、キャッシュメモリ(142)に格納される。

【0087】ステップ1211では、I/O管理リスト(503)のステータスエリア(905)にコマンド受付完了を示すフラグを設定する。このステータスエリア(905)には、コマンド終了時には実行結果が格納されるが、それまでは、ディスクアレイ制御装置(102)内でのコマンドの実行状態がイベント毎に格納され、なんらかの障害発生時にロギング情報として使用することができる。

【0088】図15は、終了処理プログラム(406)のフローチャートを示している。終了処理プログラム

(10)

特開平7-200187

17

(406)は、コマンド制御部(145)内のメモリ(404)に格納されているマイクロプログラムである。

【0089】ステップ1301では、要求されていたコマンドがREAD要求かWRITE要求かを判定する。この判定は、I/O管理リスト(503)のコマンドエリアを参照することで達成可能である。判定の結果、READ要求であればステップ1306に進み、WRITE要求であればステップ1302へ進む。

【0090】ステップ1302では、I/O管理リスト(503)のステータスエリア(905)とタグエリア(907)の内容をマージする。ステップ1303では、ステップ1302でマージしたデータを、I/O管理リスト(503)のホストエリア(903)が示すホストプロセッサ(101)へ、転送する。その後、ステップ1304に進む。WRITE処理はデータをすでにディスクドライブあるいはキャッシュエリアに書き込んだ後であるため、ホストプロセッサ(101)への情報は、処理が正常に終了したかどうかのみである。

【0091】ステップ1306は、READ処理要求の終了時に実行される。キャッシュメモリ(142)内には、要求されたデータがすでに格納されている。ステップ1306では、このデータとステータスエリア(905)とタグエリア(907)の内容とをマージする。ステップ1307では、ステップ1306でマージしたデータを、I/O管理リスト(503)のホストエリア(903)が示すホストプロセッサ(101)へ、転送する。その後、ステップ1304に進む。

【0092】ステップ1304では、I/O管理リスト(503)から、当該入出力要求を削除する。

【0093】図16は、グローバルRAID制御部(143)の構成を示している。グローバルRAID制御部(143)は、コマンドやデータの転送制御を行うアダプタ(1401)、プロセッサ(1402)、DMAC(1403)、メモリ(1404)、バススイッチ(1405)、およびパリティ生成部(1406)から構成されている。

【0094】メモリ(1404)内には、コマンド制御を行うマイクロプログラム(1407、1408、1409、1410、1411、1412)が格納されている。具体的には、I/O要求受付処理プログラム(1407)、データ回復処理プログラム(1408)、I/O要求終了処理プログラム(1409)、データ配置制御プログラム(1410)、グローバルパリティ制御プログラム(1411)、およびタグ制御プログラム(1412)である。これらの詳細は後で述べる。マイクロプログラム(1407、1408、1409、1410、1411、1412)は、プロセッサ(1402)で実行される。

【0095】外部の制御部とのデータ転送は、ダイレク

18

トメモリアクセスコントローラであるDMAC(1403)が、プロセッサ(1402)とは独立に実行する。バススイッチ(1405)は、DMAC(1403)から転送されたデータのヘッダを参照し、出力信号を振り分ける動作を行う。この動作については後で詳細に説明する。パリティ生成部(1406)は、パリティバッファ(1415)と、そのパリティバッファ(1415)に格納された(転送されてきた)データのパリティデータを作成するパリティジェネレータ(1408)とから構成される。

【0096】グローバルRAID制御部(143)の大きな制御の流れは、以下のようなものである。まず、キャッシュメモリ(142)からデータを受け取り、データ配置制御プログラム(1410)によって必要であればデータの分割を行なう。次に、分割されたデータに、各々のデータをどのローカルディスクアレイ制御装置(104、105、106、107)に転送するかを識別するためのヘッダを付加する。その後、ヘッダに従いバススイッチ1405が各ローカルディスクアレイ制御装置(104、105、106、107)に分割後のデータを転送する。

【0097】図17は、I/O要求受付処理プログラム(1407)のフローチャートを示している。

【0098】ステップ1501では、I/O管理リスト(503)のステータスエリア(905)に、グローバルRAID制御部(143)が処理要求を受け付けたことを示すフラグを、セットする。ステップ1502では、I/O管理リスト(503)のコマンドエリア(906)の内容を内部バッファに転送する。その後、ステップ1503に進み、データ配置制御プログラム(1410)に制御を移行する。

【0099】図18は、データ配置制御プログラム(1410)のフローチャートを示す。

【0100】ステップ1601では、要求処理がREADかWRITEかを判定する。もしREAD処理要求であれば、ステップ1602に進み、要求データがキャッシュメモリ(142)に存在するかどうか判定する。そうであれば、ステップ1603に進み、グローバルRAID制御部(143)がローカルディスクアレイ制御装置(104、105、106、107)に対して出力可能状態であるかどうか判定する。

【0101】ステップ1603について詳しく説明する。グローバルディスクアレイ制御装置(103)は、ローカルディスクアレイ制御装置(104、105、106、107)の上位に位置する制御装置であるため、ストライピングサイズが大きくなるために、ローカルディスクアレイ制御装置(104、105、106、107)よりもたくさんのデータをバッファリングした方がよいケースがある。例えば、4台のローカルディスクアレイ制御装置(104、105、106、107)が各

(11)

特開平7-200187

19

々4KBのデータを4台のディスクドライブに1KBずつ分割格納する構成で、グローバルディスクアレイ制御装置(103)が4台のローカルディスクアレイ制御装置(104, 105, 106, 107)にフルストライピングするケースでは、グローバルディスクアレイ制御装置(103)は、

$4KB \times 4台 = 16KB$

のデータをバッファリングすることによって4台のローカルディスクアレイ制御装置(104, 105, 106, 107)に同時に出力処理ができる。この場合、ステップ1603では、全てのローカルディスクアレイ制御装置(104, 105, 106, 107)に同時に出力可能かどうかを判定する。

【0102】しかし、このような判定が必要ないケースもある。上記の例では、フルストライピングを前提としたが、RAID4, 5のように、基本的にはデータをストライピングしない場合もある。その場合は、ホストプロセッサ(101)から転送されたデータを、バッファリングすることなく、いづれかのローカルディスクアレイ制御装置(104, 105, 106, 107)に出力処理を行なう。また、別のケースとして、仮想ドライブが複数のローカルディスクアレイ制御装置(104, 105, 106, 107)に渡って定義されていないときがある。この場合は、ローカルディスクアレイ制御装置(104, 105, 106, 107)の通常の動作通りに、ローカルディスクアレイ制御装置(104, 105, 106, 107)の分割動作に添って実行すれば良いので、グローバルディスクアレイ制御装置(103)はバッファリングする必要はない。従って、ステップ1603では、仮想ドライブ管理テーブル(502)やローカルディスクアレイ管理テーブル(504)を参照して、どのくらいのデータをバッファリングする必要があるかを判定する。

【0103】ステップ1605では、ホストコマンドのディスク要求アドレスを先頭ローカルディスクアレイ制御装置(104, 105, 106, 107)のアドレスに変換する。この処理は、ホストプロセッサ(101)が認識しているドライブと、実際にローカルディスクアレイのドライブに格納される場所とが一致していないために行なう必要がある。参照するテーブルは、仮想ドライブ管理テーブル(502)とローカルディスクアレイ管理テーブル(504)である。

【0104】例えば、ホストプロセッサ(101)からの要求ディスクアドレスが、仮想ドライブ管理テーブル(502)のVOL2の先頭から32KB目から12KBのデータ出力とする。その場合のローカルディスクアレイ制御装置(104, 105, 106, 107)の先頭アドレスの求め方は、以下の(1)～(4)の通りである。

【0105】(1) 図9の仮想ドライブ管理テーブル

20

(502)を参照して、ローカルディスクアレイ制御装置(104, 105, 106, 107)にどのように仮想ドライブが配置されているかを求める。その結果、VOL2はローカルディスクアレイ制御装置(104, 105, 106, 107)のパーティション2から構成されていることがわかる。

【0106】(2) ローカルディスクアレイ制御装置(104, 105, 106, 107)のパーティション2は、図8のローカルディスクアレイ管理テーブル(504)のカラム605, 606, 608から、それぞれ、RAIDレベル3、パリティ数1であり、パーティション2は各ローカルディスクアレイ制御装置(104, 105, 106, 107)の1GB目から始まることがわかる。

【0107】(3) 図8のローカルディスクアレイ管理テーブル(504)のカラム604, 607から、各ローカルディスクアレイ制御装置(104, 105, 106, 107)のドライブ数は5台であり、ストライピングサイズは4KBであることがわかる。パリティ数は1であることから、各ローカルディスクアレイ制御装置(104, 105, 106, 107)には、 $4KB \times (5台 - 1台) = 16KB$ 格納されることになる。

【0108】(4) その結果、要求アドレスである32KB目の物理的なアドレスは、 $32KB / 16KB + 1 = 3$ 番目のローカルディスクアレイ制御装置(106)の1GB目が先頭アドレスとなる。

【0109】次に、ステップ1605では、要求コマンドがREADかWRITEかを判定する。その結果、READであればステップ1606に進み、WRITEであればステップ1611に進む。

【0110】ステップ1606では、ローカルディスクアレイ制御装置(104, 105, 106, 107)に対して入出力要求を発行するためにコマンドを生成する。このとき、各ローカルディスクアレイ制御装置(104, 105, 106, 107)への入出力単位は、図9のカラム709のストライピングサイズを指定する。ステップ1607では、新タグを生成するためにタグ制御プログラム(1412)を実行する。タグ制御プログラム(1412)の詳細は後で述べる。

【0111】ステップ1608では、ステップ1606とステップ1607で生成したコマンドと新タグとをマージし、さらに、どのローカルディスクアレイ制御装置(104, 105, 106, 107)への要求なのかを識別するためのヘッダを付加し、バススイッチ(1405)にデータ転送を行なう。ステップ1609では、次のローカルディスクアレイ制御装置を選択する。この操作は、上記物理アドレスを求める際に行なった計算と同じ要領で可能である。単に格納要求アドレスに図9のカラム709のストライピングサイズを加えたアドレスとして再計算するだけで可能となる。

(12)

特開平7-200187

21

【0112】次に、ステップ1610では、全てのデータを処理し終わったかどうかを判定し、まだ処理できていなければステップ1606から繰り返し、処理し終われば当処理を終了する。

【0113】ステップ1605でWRITEと判定されたときは、ステップ1611に進む。ステップ1611は、グローバルパリティ制御プログラムの実行を意味している。この処理についてはあとで詳細説明を行なう。

【0114】ステップ1612では、ローカルディスクアレイ制御装置(104, 105, 106, 107)に対して入出力要求を発行するためにコマンドを生成する。このとき、各ローカルディスクアレイ制御装置(104, 105, 106, 107)への入出力単位は、図9のカラム709のストライピングサイズを指定する。ステップ1613では、新タグを生成するためにタグ制御プログラム(1412)を実行する。タグ制御プログラム(1412)の詳細は後で述べる。

【0115】次に、ステップ1614では、ステップ1608と同様に、ローカルディスクアレイ制御装置(104, 105, 106, 107)への入出力要求を発行する。ステップ1615はステップ1609と同等の処理であり、ステップ1616はステップ1610と同等の処理である。

【0116】図19は、パススイッチ(1405)の動作を説明したものである。

【0117】この例では、4つのコマンド(1709, 1710, 1711, 1712)がセクタ(1414)によって分散転送される様子を示している。各コマンドは、READ/WRITEなどのコマンドフィールド(1701, 1703, 1705, 1707)と、転送先を示すバス番号(1702, 1704, 1706, 1708)とから構成されている。

【0118】例えば、バケット1712はバス番号が1、バケット1711はバス番号が2、バケット1710はバス番号が3、バケット1709はバス番号が4である。セクタ(1414)は、このバス番号を参照することによって、バス番号に続くデータをどのバスに転送すべきなのかを判定して、バスのスイッチングを行なう。例えば、バス番号が1であれば、セクタ(1414)はバス番号1を選択し、データを転送する(1713)。同様に、バケット1711, 1710, 1709も、セクタ(1414)によって、1714, 1715, 1716に分散して転送される。

【0119】図20は、グローバルパリティ制御プログラム(1411)のフローチャートを示したものである。

【0120】ステップ1801では、仮想ドライブ管理テーブル(502)からパリティ位置の算出を行なう。例えば、要求ディスクアドレスが仮想ドライブ管理テーブル(502)のVOL2の先頭から32KB目から1

22

2KBのデータ出力とする。その場合のパリティ位置の求め方は以下の(1)～(4)の通りである。

【0121】(1)仮想ドライブ管理テーブル(502)を参照して、ローカルディスクアレイ制御装置(104, 105, 106, 107)にどのように仮想ドライブが配置されているかを求める。その結果、VOL2はローカルディスクアレイ制御装置(104, 105, 106, 107)のパーティション2から構成されていることがわかる。

【0122】(2)ローカルディスクアレイ制御装置(104, 105, 106, 107)のパーティション2は、ローカルディスクアレイ管理テーブル(504)のカラム605, 606, 608から、それぞれ、RAIDレベル3、パリティ数1であり、パーティション2は各ローカルディスクアレイ制御装置(104, 105, 106, 107)の1GB目から始まることかわかる。

【0123】(3)グローバルRAIDレベルは3であることから、パリティは割当最終ローカルディスクアレイ(107)であることがわかる。

【0124】(4)その結果、パリティの物理的なアドレスは、ローカルディスクアレイ(107)の1GB目からとなる。

【0125】上記パリティ位置計算は一例であり、ローカルディスクアレイ制御装置(104, 105, 106, 107)のRAIDレベルや、グローバルディスクアレイ制御装置(103)のRAIDレベルによって異なってくるが、各RAIDレベルの配置規則を前述の演算に適用することで容易に実現可能である。

【0126】次に、ステップ1802では、パリティ演算に現在ドライブに格納されているデータが必要かどうかの判定を行なう。つまり、RAIDレベルによっては、パリティ全体を更新できないことがある。例えばRAIDレベル5等である。パリティは、RAIDレベル5に限らず複数のデータから一つあるいは一つ以上のパリティを生成する。このとき、あるパリティに関連する全てのデータが同時に更新されれば、その新しいすべてのデータから全く新しいパリティを生成できる。しかし、RAIDレベル5等のように部分的なデータの更新である場合は、まず更新すべき箇所の過去のデータの情報をパリティから削除した後、新しいデータにおけるパリティを作成しなければならない。

【0127】ステップ1802の判定結果から旧データの必要がない場合はステップ1806に進み、そうでなければステップ1803に進む。ステップ1803では、旧データの読み込みのために、ローカルディスクアレイ制御装置(104, 105, 106, 107)に対する入力コマンドを生成する。この処理は、図18のステップ1606などと同等である。入力アドレスは、ステップ1801で求めたアドレスを使用する。ステップ

(13)

特開平7-200187

23

1804では、新タグを生成するためにタグ制御プログラム(1412)を実行する。この処理については、後で詳細に説明する。ステップ1805では、ローカルディスクアレイ制御装置(104, 105, 106, 107)に対して入力要求を発行する。この処理は図18のステップ1608等と同等である。

【0128】ステップ1806では、キャッシュ領域からパリティ生成部(1406)へ、グローバルディスクアレイ制御装置(103)のストライピングサイズを指定して、データ転送を行なう。必要であればステップ1805で取得した旧データも同時に転送する。グローバルディスクアレイ制御装置(103)のストライピングサイズは、図9のカラム709のストライピングサイズを参照することで求めることができる。この動作については後で詳細に説明する。

【0129】次に、ステップ1807では、パリティ生成部(1406)から、生成されたパリティを読み取る。ステップ1808では、ステップ1801で計算したパリティ位置に対して、ステップ1807で取得した新しいパリティを書き込むためのコマンドを生成する。ステップ1809では、ステップ1808で生成したコマンドに対するタグを取得するため、タグ制御プログラム(1412)を実行する。ステップ1810では、ステップ1801で求めたパリティ位置に新しいパリティを書き込むために出力処理を行なう。

【0130】図21は、グローバルパリティの生成の様子を示している。

【0131】1901はパリティを生成するデータ群であり、1902はグローバルディスクアレイ制御装置(103)のストライピングサイズである。この2つのデータをパリティ生成部(1406)に転送することで、まずデータ群(1901)をグローバルディスクアレイ制御装置(103)のストライピングサイズに分割し(1903, 1904, 1905)、それぞれについて排他的論理和を取る(1906, 1907)。これにより、グローバルパリティが生成される。

【0132】図22は、ホストプロセッサ(101)から転送された入出力コマンドが、ディスク装置(133, 134, 135, 136)に格納あるいは読みだしされるまでに、どのように変化するかを示している。

【0133】ホストプロセッサ(101)からは、データ(2001)、データ長(2002)、コマンド(2003)、ホスト識別子(2004)、タグ(2005)、および仮想ドライブ識別子(2006)が、一つのケットとなって転送される。それを受けたグローバルディスクアレイ制御装置(103)は、ローカルディスクアレイ制御装置(104, 105, 106, 107)に対して分割して入出力要求を発行するために、複数のコマンドを生成する。

【0134】その一つ一つは、2007-2012で示

24

されるように、データ、データ長、コマンド、ホスト識別子、タグ、およびローカルディスクアレイ識別子(バス番号)が一つのケットとなっている。ここで、ケットの形式は、2001-2006と同じであるが、内容は異なる。

【0135】例えば、データ(2007)は分割されているためにデータ(2001)に比べて少ない。それに伴い、データ長(2008)もデータ長(2002)より短い値がセットされている。コマンド(2009)は、格納あるいは読みだし位置がローカルディスクアレイのアドレスに変わる。ホスト識別子(2010)は、ホストプロセッサ(101)の識別子とは異なり、この場合はグローバルディスクアレイ制御装置(103)の識別子である。タグ(2011)は、ホストプロセッサ(101)から転送されたタグではなく、新タグである。この理由は既に述べている。2012は、ホストプロセッサ(101)からは見えないローカルディスクアレイ識別子(バス番号)が格納されている。

【0136】ローカルディスクアレイ制御装置(104, 105, 106, 107)内では、2013-2017に示すように、データ、データ長、コマンド、ホスト識別子、およびドライブ番号が一つのケットとなってドライブに転送される。前述のグローバルディスクアレイ制御装置(103)内でのコマンドの変化と同じように、分割が行なわれる。

【0137】図23は、タグ制御プログラム(1412)のフローチャートを示している。このタグは、前述の通り、ホストプロセッサ(101)から転送されたタグを、グローバルディスクアレイ制御装置(103)からローカルディスクアレイ制御装置(104, 105, 106, 107)に対する入出力には使用できないことから、タグの変換が必要であることによる。

【0138】ステップ2101では、入出力を行なうローカルディスクアレイ制御装置(104, 105, 106, 107)に対応する、コマンドキューリスト(506)中のフリータグポインタ(1102, 1104, 1106, 1108)を検索する。フリータグポインタは、ローカルディスクアレイ制御装置(104, 105, 106, 107)毎に設けられており、各ローカルディスクアレイ制御装置(104, 105, 106, 107)に対する入出力に使用されていないタグがリスト形式でチェインされている(図13)。ステップ2102では、このリストの中からタグを一つ取得する。

【0139】次に、ステップ2103では、取得したタグを使用中リストに繋ぎ換える。これにより、取得したタグは他の要求により使用されることが無くなり、同一ローカルディスクアレイ制御装置(104, 105, 106, 107)内で重複するタグが使用されることはない。ステップ2104では、取得したタグを、I/O管理リスト(503)中の新タグエリア(908)に格納

(14)

特開平7-200187

25

する。これは、ホストプロセッサ（101）から転送されたタグと関連づけるために行なう。これにより、入出力完了時に、ホストプロセッサ（101）に対してホストプロセッサ（101）から転送されたタグを返送するとき、容易に検索することが可能となる。ステップ2105では、タグをコマンドに付加し、コマンドキューリスト（506）内のコマンド格納エリアに格納する。

【0140】図24は、データ回復制御プログラム（1408）のフローチャートを示している。

【0141】ディスクアレイ装置にはデータの回復手段が設けられている。本発明では、通常のディスクアレイに比べより高信頼のディスクアレイにするために、パリティの階層化を行なっている。ローカルディスクアレイ制御装置（104、105、106、107）が管理するパリティと、グローバルディスクアレイ制御装置（103）が管理するグローバルパリティである。グローバルパリティは、複数のローカルディスクアレイ制御装置（104、105、106、107）に渡るパリティである。

【0142】ローカルディスクアレイ制御装置（104、105、106、107）が管理するパリティは、そのローカルディスクアレイ制御装置（104、105、106、107）内でのみ使用することができる。従って、ローカルディスクアレイ制御装置（104、105、106、107）内のいずれかのディスクドライブに障害が発生した場合に、そのパリティを使用して、障害ドライブのデータを回復することができる。

【0143】グローバルパリティは、複数のローカルディスクアレイ制御装置（104、105、106、107）に渡って作成されているため、前述のディスクドライブ単体の障害時にも使用することは可能であるが、それ以上にローカルディスクアレイ制御装置（104、105、106、107）そのものが障害を起こしたときに使用することができることの意味が大きい。

【0144】図24のフローチャートでは障害回復を階層化した動作を示している。

【0145】ステップ2201では、ローカルディスクアレイ制御装置（104、105、106、107）の回復処理が可能かどうか判定する。この判定にはさまざまな要因が考えられる。例えば、障害回復機構の障害や、パリティで回復できる範囲を越えた障害時には、ローカルディスクアレイ制御装置（104、105、106、107）単体では障害回復ができない。この場合は、ステップ2202に進む。

【0146】ステップ2202では、パリティ領域の障害かどうか判定する。パリティ障害であれば、データ部分の障害ではないため、ここでは回復処理を行なわない。ステップ2203では、障害を起こしたローカルディスクアレイ以外のデータを取得する。この場合、パリティデータも含む。ステップ2204では、ステップ2

26

203で取得したデータあるいはパリティから、障害を起こした部分のデータを回復する。

【0147】なお、ここでは述べていないが、グローバルパリティのパリティも存在することがある。これは、グローバルディスクアレイ制御装置（103）がローカルディスクアレイ制御装置（104、105、106、107）にグローバルパリティデータを格納したときに、ローカルディスクアレイ制御装置（104、105、106、107）内では、グローバルパリティデータも通常のデータと同じくパリティを作成／格納するためである。本発明では、このようにパリティ情報も階層化されているため、高い信頼性を得ることができる。

【0148】図25は、I/O要求終了処理プログラム（1409）のフローチャートを示している。

【0149】ステップ2301では、コマンドキューリスト（506）から終了する要求に対応するリストをサーチする。ステップ2302では、I/O管理リスト（503）からステップ2301と同様にサーチする。ステップ2304では、I/O管理リスト（503）のステータスエリア（905）に、ローカルディスクアレイ制御装置（104、105、106、107）から転送された終了状態を、格納する。

【0150】ステップ2304では、コマンドキューリスト（506）からステップ2301でサーチしたリストを削除し、未使用リストへ繋ぎ換える。これにより、本処理で使用したタグは、他の入出力のために使用可能となる。ステップ2305では、ホストプロセッサ（101）から転送されたタグをI/O管理リスト（503）のタグエリア（907）から取り出し、コマンド制御部（145）に通知することで処理を終了する。

【0151】次に、ローカルディスクアレイ制御装置（104、105、106、107）の詳細を述べる。この中で、コマンド制御部（113、114、115、116）は、グローバルディスクアレイ制御装置103のコマンド制御部141と同等である。また、インタフェース制御部（109、110、111、112）は、グローバルディスクアレイ制御装置103のインタフェース制御部108と同等である。また、インタフェースコントローラ（129、130、131、132）は、グローバルディスクアレイ制御装置103のインタフェースコントローラ144と同等である。

【0152】図26は、ローカルディスクアレイ制御装置（104、105、106、107）のRAID制御部（125、126、127、128）の構成を示している。なお、図26において、図16に示したグローバルRAID制御部143の各部と共通の部分は、同じ番号で示している。

【0153】RAID制御部（125、126、127、128）は、コマンドやデータの転送制御を行うアダプタ（1401）、プロセッサ（1402）、DMA

(15)

特開平7-200187

27

C (1403)、メモリ (1404)、バススイッチ (1405)、およびパリティ生成部 (1406) から構成されている。

【0154】メモリ (1404) 内には、コマンド制御を行うマイクロプログラム (1407、2404、2401、2402、2403) が格納されている。具体的には、I/O要求受付処理プログラム (1407)、データ回復処理プログラム (2404)、I/O要求終了処理プログラム (2401)、データ配置制御プログラム (2403)、およびパリティ制御プログラム (2402) である。

【0155】I/O要求受付処理プログラム (1407) は、図17と同様であるので説明を省略する。I/O要求終了処理プログラム (2401)、データ配置制御プログラム (2403)、パリティ制御プログラム (2402)、およびデータ回復処理プログラム (2404) の詳細は後で述べる。マイクロプログラム (1407、2404、2401、2402、2403) はプロセッサ (1402) で実行される。

【0156】外部の制御部とのデータ転送は、DMAC (1403) がプロセッサ (1402) とは独立に実行する。バススイッチ (1405) は、DMAC (1403) から転送されたデータのヘッダを参照し、出力信号を振り分ける動作を行う。パリティ生成部 (1606) は、パリティバッファ (1415) と、そのパリティバッファ (1415) に格納された (転送されてきた) データのパリティデータを作成するパリティジェネレータ (1408) とから構成される。

【0157】RAID制御部 (125、126、127、128) の大きな制御の流れは、以下のようなものである。まず、グローバルディスクアレイ制御装置 (103) からデータを受け取り、データ配置制御プログラム (2403) によって必要であればデータの分割を行なう。次に、分割されたデータに、各々のデータをどのディスクドライブ (137、138、139、140) に転送するかを識別するためのヘッダを付加する。その後、ヘッダに従いバススイッチ (1405) が各ディスクドライブ (137、138、139、140) に分割後のデータを転送する。

【0158】図27は、制御メモリ (117、118、119、120) に格納されているリストを示している。2502はI/O管理リスト、2503はキャッシュ管理リストである。

【0159】I/O管理リスト (2502) は、グローバルディスクアレイ制御装置 (103) から受け取った入出力要求を入出力が環境するまで入出力状態等と共に管理するテーブルである。キャッシュ管理リスト (2503) は、キャッシュメモリ (121、122、123、124) 内に格納されているデータとそのアドレスが格納されている。I/O管理リスト (2502) とキ

28

ャッシュ管理リスト (2503) は、図7のI/O管理リスト503 (図11) とキャッシュ管理リスト505 (図12) と同等のリストである。

【0160】図28は、I/O要求終了処理プログラム (2401) のフローチャートを示している。

【0161】ステップ2601では、I/O管理リスト (2502) から終了する入出力要求のリストをサーチする。ステップ2602では、I/O管理リスト (2502) のステータスエリアにディスク装置 (133、134、135、136) から転送された終了状態を格納することで処理を終了する。

【0162】図29は、データ配置制御プログラム (2403) のフローチャートを示す。

【0163】ステップ2701では、要求処理がREADかWRITEかを判定する。もし、READ処理要求であれば、ステップ2702に進み、要求データがキャッシュメモリ (142) に存在するかどうか判定する。ステップ2701でWRITE処理要求のときは、ステップ2703に進み、RAID制御部 (125、126、127、128) がディスク装置 (133、134、135、136) に対して出力可能状態であるかどうか判定する。

【0164】ステップ2704では、グローバルディスクアレイ制御装置 (103) のディスク要求アドレスを先頭ディスク装置 (133、134、135、136) のアドレスに変換する。ステップ2705では、要求コマンドがREADかWRITEかを判定する。その結果、READであればステップ2706に進み、WRITEであればステップ2709に進む。

【0165】ステップ2706では、ディスクドライブ (137、138、139、140) に対して入出力要求を発行する。ステップ2707では、次のディスクドライブ (137、138、139、140) を選択する。ステップ2708では全てのデータ処理が終了したかどうか判定し、終了していなければステップ2706から繰り返す。

【0166】ステップ2709はステップ2706と同等である。ステップ2710はステップ2707と同等である。ステップ2711はステップ2708と同等である。ステップ2712では、パリティ制御プログラム (2402) を実行する。

【0167】図30は、パリティ制御プログラム (2402) のフローチャートを示したものである。

【0168】ステップ2801では、RAIDレベル、ドライブ数、および要求ディスクアドレスからパリティ位置の算出を行なう。ステップ2802では、パリティ更新のために旧データが必要かどうか判定する。ステップ2802の判定結果から、旧データの必要がない場合はステップ2804に進み、そうでなければステップ2803に進む。

(16)

特開平7-200187

29

【0169】ステップ2803では、旧データの読み込みのために、ディスクドライブ(137, 138, 139, 140)に対して旧データの入力要求を発行する。入力アドレスは、ステップ2801で求めたアドレスを使用する。

【0170】ステップ2804では、キャッシュ領域からパリティ生成部(1406)へデータ転送を行なう。必要であればステップ2803で取得した旧データも同時に転送する。ステップ2805では、パリティ生成部(1406)から、生成されたパリティを読み取る。ステップ2806では、ステップ2801で求めたパリティ位置に新しいパリティを書き込むために出力処理を行なう。

【0171】図31は、データ回復制御プログラム(2404)のフローチャートを示している。ディスクアレイ装置にはデータの回復手段が設けられている。従って、ローカルディスクアレイ制御装置(104, 105, 106, 107)内のいずれかのディスクドライブに障害が発生した場合に、そのパリティを使用して、障害ドライブのデータを回復することができる。

【0172】まず、ステップ2901では、パリティ領域の障害かどうか判定する。パリティ障害であれば、データ部分の障害ではないため、ここでは回復処理を行なわない。ステップ2902では、障害を起こしたディスクドライブ以外のデータを取得する。この場合、パリティデータも含む。ステップ2903では、ステップ2902で取得したデータあるいはパリティから、障害を起こした部分のデータを回復する。

【0173】図32は、コマンド受付処理プログラムのフローチャートを示している。コマンド受付処理プログラムは、コマンド制御部(113, 114, 115, 116)内のメモリに格納されているマイクロプログラムである。

【0174】ステップ3001では、キャッシュ確保可能かどうかをチェックする。もし可能であればステップ3002へ移り、そうでなければステップ3013へ移りコマンド異常終了処理を行う。ステップ3002では、キャッシュメモリ(121, 122, 123, 124)の所要分領域を確保する。ステップ3003では、キャッシュ管理リスト(2503)のフリーポインタ(1001)でポイントされているリストから、使用中ポインタ(1002)がポイントしているリスト内に入れる。

【0175】ステップ3004では、I/O管理リスト(2502)中に空きリストが存在するかどうか調べる。もし、空きリストが存在すればステップ3005に進み、そうでなければステップ3013に進み異常終了処理を行う。

【0176】ステップ3005では、新しいI/O管理リスト(2502)を追加する。ステップ3006で

30

は、ホストIDをI/O管理リスト(2502)のエリア(903)に格納する。ステップ3008では、コマンド受付時間をI/O管理リスト(2502)のエリア(904)に格納する。この時間は、もしディスクアレイ制御装置(102)に障害が発生した場合などの保守に使用することができる。

【0177】ステップ3009では、コマンドをI/O管理リスト(2502)のエリア(906)に格納する。ステップ3010では、ステップ3002で取得したキャッシュメモリ(121, 122, 123, 124)領域にコマンドを転送する。ここでコマンドとはREAD/WRITEなどのディスク装置に対する指示と、WRITE要求であればデータも同様にキャッシュメモリ(142)に格納される。

【0178】ステップ3011では、I/O管理リスト(2502)のステータスエリア(905)にコマンド受付完了を示すフラグを設定する。このステータスエリア(905)は、コマンド終了時には実行結果が格納されるが、それまでは、ローカルディスクアレイ制御装置(104, 105, 106, 107)内でのコマンドの実行状態がイベント毎に格納され、なんらかの障害発生時にロギング情報として使用することができる。

【0179】図33は、終了処理プログラムのフローチャートを示している。終了処理プログラムは、コマンド制御部(113, 114, 115, 116)内のメモリに格納されているマイクロプログラムである。

【0180】ステップ1301では、要求されていたコマンドがREAD要求かWRITE要求かを判定する。この判定は、I/O管理リスト(2502)のコマンドエリアを参照することで達成可能である。判定の結果、READ要求であればステップ3103に進み、WRITE要求であればステップ3102へ進む。

【0181】ステップ3102では、I/O管理リスト(2502)のステータスエリア(905)をグローバルディスクアレイ制御装置(103)へ転送する。その後、ステップ3105に進む。

【0182】ステップ3103は、READ処理要求の終了時に実行される。キャッシュメモリ(121, 122, 123, 124)内には、要求されたデータがすでに格納されている。ステップ3103では、このデータとステータスエリア(905)の内容をマージする。ステップ3104では、ステップ3103でマージしたデータをI/O管理リスト(2502)のホストエリア(903)が示すグローバルディスクアレイ制御装置(103)へ転送する。その後、ステップ3105に進む。

【0183】ステップ3105では、I/O管理リスト(2502)から、当該入出力要求を削除する。

【0184】図34は、本実施例における障害回復動作の一例を示す。この図を参照して、本発明による障害回

(17)

特開平7-200187

31

復動作の一例を説明する。

【0185】本例では、ローカルディスクアレイ制御部（104）が管理するディスクドライブの一つに障害が発生し、さらに、ローカルディスクアレイ制御装置（105）に障害が発生したとき、ホストプロセッサ（101）からデータの読みだし要求が出されたときの回復動作を示している。

【0186】まず、ローカルディスクアレイ制御装置（104）は、障害が発生したドライブ以外のドライブからデータを読みだし、障害が発生したドライブのデータを修復する（3204）。修復されたデータと正常ドライブのデータはマージされて（3203）、グローバルディスクアレイ制御装置（103）へ転送する。

【0187】ローカルディスクアレイ制御装置（105）は、要求されたデータ（3205）をそのままグローバルディスクアレイ制御装置（103）へ転送する。グローバルディスクアレイ制御装置（103）は、ローカルディスクアレイ制御装置（105）が単独で修復不可能であることを判断し、グローバルバリティを使用したデータ修復を行なう。

【0188】そのために、グローバルバリティが格納されているローカルディスクアレイ制御装置（107）から、回復に必要なグローバルバリティ（3206）を読み出す。グローバルバリティ（3206）と正常にデータが読みだされたローカルディスクアレイのデータは、グローバルディスクアレイ制御装置（103）内で、障害が発生したローカルディスクアレイ制御装置（105）のデータを修復する（3202）。修復されたデータは、正常にデータが読みだされたローカルディスクアレイのデータとマージされ（3201）、ホストプロセッサ（101）へ転送される。

【0189】このように、本発明では、ドライブ障害だけでなく、ディスクアレイそのものに障害が発生しても、ホストプロセッサ（101）の処理を中断することがない。

【0190】図35は、本発明によるディスクアレイ装置のバリティデータの格納の一例を示している。図中、Pはローカルディスクアレイ制御装置（104、105、106、107）により生成されたバリティを示し、GPはグローバルディスクアレイ制御装置（103）により生成されたバリティを示す。

【0191】3301-3304は、ローカルディスクアレイ制御装置（104、105、106、107）のRAIDレベルが3または4のときのバリティを示す。それに対し、3300はグローバルディスクアレイ制御装置（103）が生成したバリティであり、このときグローバルディスクアレイ制御装置（103）のRAIDレベルは3または4である。

【0192】同様に、3309-3312は、ローカルディスクアレイ制御装置（104、105、106、1

32

07）のRAIDレベルが3または4のときのバリティを示す。3305-3308はグローバルディスクアレイ制御装置（103）が生成したバリティであり、このときグローバルディスクアレイ制御装置（103）のRAIDレベルは5である。このように、データのみならずバリティも複数のローカルディスクアレイ制御装置（104、105、106、107）に渡って格納される。

【0193】3313のグローバルバリティも同様である。このとき、ローカルディスクアレイ制御装置（104、105、106、107）はRAIDレベルが5である。

【0194】図36は、図35と同様にバリティデータの格納の一例を示している。図35では、グローバルディスクアレイ制御装置（103）とローカルディスクアレイ制御装置（104、105、106、107）の両方でバリティを生成／格納することで、非常に高い信頼性を得られる例を示した。図36は、信頼性よりも大容量を優先する使用方法を示した例である。

【0195】3401は、当ディスクアレイ装置の中で唯一つのバリティである。他のローカルディスクアレイ制御装置（104、105、106）にはバリティは存在しない。こうすることで、容量を優先するディスクアレイを構成することができる。

【0196】これと同様の別の例として、ローカルディスクアレイ制御装置（107）のすべてのディスクドライブにグローバルバリティを格納することもできる。3402は、RAID5のグローバルバリティである。しかし、他のローカルディスクアレイ制御装置（104、105、106）にはバリティは存在しない。

【0197】同じRAID5のグローバルバリティでも、3403-3406のようにローカルディスクアレイ制御装置（104、105、106、107）に分散して配置することも可能である。

【0198】[実施例2]次に、本発明の第2の実施例を説明する。

【0199】図37は、本発明の第2の実施例の全体構成図を示したものである。図1と共通の部分は同じ番号で示した。図1と大きく異なるところは、グローバルディスクアレイ制御装置（3513）とローカルディスクアレイ制御装置（3502、3503、3504、3405）との接続に、共通バス（3506）を用いたところである。制御メモリ（3511）、グローバルRAID制御部（3501）、およびコマンド制御部（3507-3510）については後で詳細説明を行なうが、その他の部分については図1と同等であるので説明を省略する。

【0200】図38は、グローバルRAID制御部（3501）の構成を示している。基本的な構成は、図16と同じであるが、マイクロプログラム（3601）とマ

(18)

特開平7-200187

33

ルチキャスト制御部(3602)が異なる。

【0201】グローバルRAID制御部(3501)は、ホストプロセッサ(101)から入出力要求を受け取り、データをローカルディスクアレイ制御装置(3502, 3503, 3504, 3405)に対して転送する。このとき、グローバルRAID制御部(3501)は、マルチキャスト制御部(3602)により共通バス(3506)に接続された全てのローカルディスクアレイ制御装置(3502, 3503, 3504, 3405)を転送先として同一のコマンドを転送する。

【0202】図39は、ローカルディスクアレイ制御装置(3502, 3503, 3504, 3405)内の制御メモリ(3511)の構造を示している。

【0203】テーブルヤリスト(502, 503, 504, 505)は、図7と同等であるが、ローカルディスクアレイ識別子(3702)が追加されている。ローカルディスクアレイ識別子(3702)は、ローカルディスクアレイ制御装置(3502, 3503, 3504, 3405)毎に異なる識別子が設定される。ローカルディスクアレイ識別子(3702)は、自ローカルディスクアレイ制御装置(3502, 3503, 3504, 3405)がグローバルRAID制御部(3501)から転送されたコマンドやデータを実行すべきかどうかを判断するために使用する。

【0204】図40は、コマンド受付処理プログラムのフローチャートを示している。コマンド受付処理プログラムは、コマンド制御部(3507, 3508, 3509, 3510)内のメモリに格納されているマイクロプログラムである。

【0205】ステップ3801では、制御メモリ(3511)内のローカルディスクアレイ識別子(3702)と転送されたコマンドとから、自ローカルディスクアレイが実行すべき内容かどうかを判定する。この判定は、ローカルディスクアレイ識別子(3702)と仮想ドライブ管理テーブル(502)を用いて行なわれる。例えば、グローバルRAID制御部(3501)から仮想ドライブnのブロックmに対する入出力要求があった場合は以下のように判定する。

【0206】(1)仮想ドライブ管理テーブル(502)から仮想ドライブn、ブロックmが格納されているローカルディスクアレイの識別子を得る。

(2)得られた識別子と制御メモリ(3511)内のローカルディスクアレイ識別子(3702)とを比較し、自ローカルディスクアレイが実行すべきかどうか判定する。

【0207】以上の動作により、ステップ3801の判定が可能となる。自ローカルディスクアレイで実行すべきコマンドであったときはステップ3802に進み、総出内ときは処理を終了する。

【0208】ステップ3802では、キャッシュ確保可

34

能かどうかをチェックする。もし可能であればステップ3803へ移り、そうでなければステップ3814へ移りコマンド異常終了処理を行う。ステップ3803では、キャッシュメモリ(121-124)の所要分領域を確保する。ステップ3804では、キャッシュ管理リスト(505)のフリーポインタ(1001)でポイントされているリストから、所要分のリストを、使用中ポインタ(1002)がポイントしているリスト内に入れる。

【0209】次に、ステップ3805では、I/O管理リスト(503)中に空きリストが存在するかどうか調べる。もし、空きリストが存在すればステップ3806に進み、そうでなければステップ3814に進み異常終了処理を行う。

【0210】ステップ3806では、I/O管理リスト(503)に新しいリストを追加する。ステップ3807では、ホストIDをI/O管理リスト(503)のエリア(903)に格納する。ステップ3809では、コマンド受付時間をI/O管理リスト(503)のエリア(904)に格納する。この時間は、もしディスクアレイ制御装置(102)に障害が発生した場合などの保守に使用することができる。ステップ3810では、コマンドをI/O管理リスト(503)のエリア(906)に格納する。

【0211】ステップ3811では、ステップ3803で取得したキャッシュメモリ(121-124)領域にコマンドを転送する。ここでコマンドとはREAD/WRITEなどのディスク装置に対する指示と、WRITE要求であればデータも同様にキャッシュメモリ(121-124)に格納される。ステップ3812では、I/O管理リスト(503)のステータスエリア(905)にコマンド受付完了を示すフラグを設定する。このステータスエリア(905)には、コマンド終了時には実行結果が格納されるが、それまでは、ディスクアレイ制御装置(102)内でのコマンドの実行状態がイベント毎に格納され、なんらかの障害発生時にロギング情報として使用することができる。

【0212】図41は、データ配置制御プログラム(1410)の動作フローチャートを示している。

【0213】ステップ3901では、要求処理がREADかWRITEかを判定する。もし、READ処理要求であれば、ステップ1602に進み、要求データがキャッシュメモリ(142)に存在するかどうか判定する。ステップ3901でWRITE処理要求であれば、ステップ3903に進む。ステップ3903では、ホストプロセッサ(101)からの要求をそのままローカルディスクアレイ制御装置(3502, 3503, 3504, 3405)に転送する。

【0214】図42は、本実施例2の動作の一例を示している。

(19)

特開平 7-200187

35

【0215】グローバルディスクアレイ制御装置(3513)から、データ4002、入出力データ長4003、入出力先頭アドレス4004、仮想ドライブ識別子4005、およびコマンド4006が、一つのバケットとして、全てのローカルディスクアレイ制御装置(3502, 3503, 3504, 3405)に転送される(4001)。

【0216】ローカルディスクアレイ制御装置(3502, 3503, 3504, 3405)での実行判定の結果、3502, 3505では自ローカルディスクアレイ制御装置には関係のない入出力となりコマンドキャンセルを行なう。ローカルディスクアレイ制御装置(3503, 3504)は、自ローカルディスクアレイ内の領域であると判断し、グローバルディスクアレイ制御装置(3513)から転送されたコマンドを解釈し実行する(4007, 4008)。

【0217】こういった処理により、ローカルディスクアレイ制御装置(3502, 3503, 3504, 3405)の接続性を良くすることが可能となる。つまり、グローバルディスクアレイ制御装置(3513)内では、入出力制御においてローカルディスクアレイ制御装置(3502, 3503, 3504, 3405)を意識する必要がなくなった。従って、ローカルディスクアレイ制御装置を新たに追加しても、グローバルディスクアレイ制御装置(3513)の変更は必要ない。さらには、ローカルディスクアレイ制御装置も、制御メモリ(3507-3510)内のローカルディスクアレイ識別子を変更するだけで増設が可能となる。

【0218】

【発明の効果】本発明によれば、多数のディスク装置を接続しても、高速な入出力が可能であり、コストを上げることがなく、またディスク管理も容易で信頼性を低下させることもない。

【図面の簡単な説明】

【図1】本発明によるディスクアレイの一実施例の全体図を示す。

【図2】本発明による実施例1の動作概要を示す。

【図3】本発明による実施例1の動作概要を示す。

【図4】本発明による実施例1の動作概要を示す。

【図5】本発明による実施例1の動作概要を示す。

【図6】コマンド制御部の構成を示す。

【図7】制御メモリの構成を示す。

【図8】ローカルディスクアレイ管理テーブルの構成を示す。

【図9】仮想ドライブ管理テーブルの構成を示す。

【図10】仮想ドライブの概要を示す。

【図11】I/O管理リストの構成を示す。

【図12】キャッシュ管理リストの構成を示す。

【図13】コマンドキューリストの構成を示す。

【図14】コマンド受付処理プログラムのフローを示す。

36

す。

【図15】終了処理プログラムのフローを示す。

【図16】グローバルRAID制御部の構成を示す。

【図17】I/O要求受付処理プログラムのフローを示す。

【図18】データ配置制御プログラムのフローを示す。

【図19】バススイッチの動作を示す。

【図20】グローバルパリティ制御プログラムのフローを示す。

【図21】グローバルパリティの生成概要を示す。

【図22】コマンドバケットの流れを示す。

【図23】タグ制御プログラムのフローを示す。

【図24】データ回復制御プログラムのフローを示す。

【図25】I/O要求終了プログラムのフローを示す。

【図26】ローカルディスクアレイのRAID制御部の構成を示す。

【図27】ローカルディスクアレイの制御メモリの構成を示す。

【図28】ローカルディスクアレイのI/O要求終了プログラムのフローを示す。

【図29】ローカルディスクアレイのデータ配置制御プログラムのフローを示す。

【図30】ローカルディスクアレイのパリティ制御プログラムのフローを示す。

【図31】ローカルディスクアレイのデータ回復制御プログラムのフローを示す。

【図32】ローカルディスクアレイのコマンド受付処理プログラムのフローを示す。

【図33】ローカルディスクアレイの終了処理プログラムのフローを示す。

【図34】本発明によるデータ回復処理の動作を示す。

【図35】本発明によるデータ格納例を示す。

【図36】本発明によるデータ格納例を示す。

【図37】本発明による実施例2の全体構成図を示す。

【図38】実施例2のグローバルRAID制御部の構成を示す。

【図39】実施例2の制御メモリの構成を示す。

【図40】実施例2のコマンド受付処理プログラムのフローを示す。

【図41】実施例2のデータ配置制御プログラムのフローを示す。

【図42】実施例2の動作例を示す。

【符号の説明】

101…ホストプロセッサ、102…ディスクアレイ制御装置、108…インタフェース制御部、133, 134, 135, 136…ディスク装置、137, 138, 139, 140…ディスクドライブ、103…グローバルディスクアレイ制御装置、104, 105, 106, 107…ローカルディスクアレイ制御装置、141…コマンド制御部、145…制御メモリ、142…キャッシュ

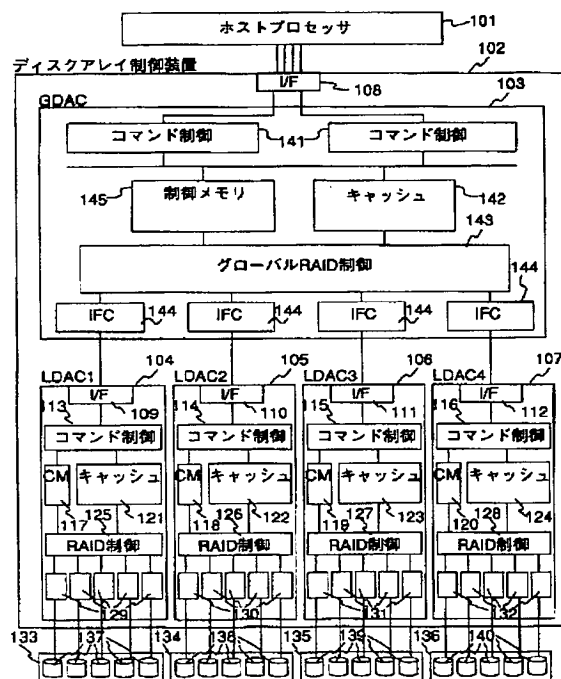
(20)

特開平7-200187

37

メモリ、143…グローバルRAID制御部、144…インタフェースコントローラ、402…プロセッサ、502…仮想ドライブ管理テーブル、503…I/O管理リスト、504…ローカルディスクアレイ管理テーブル、505…キャッシュ管理リスト、506…コマンドキューリスト、405…コマンド受付処理プロ

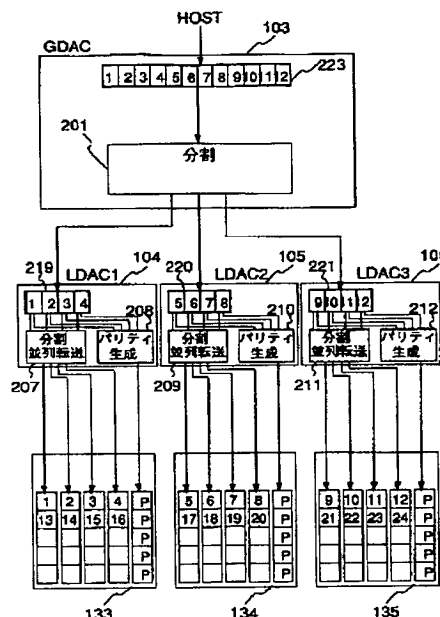
【図1】



38

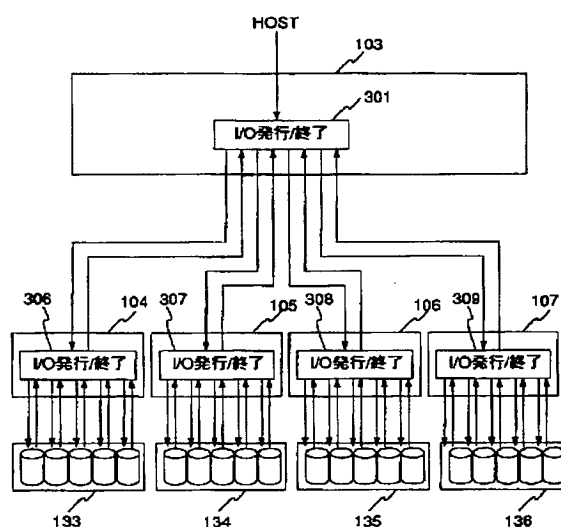
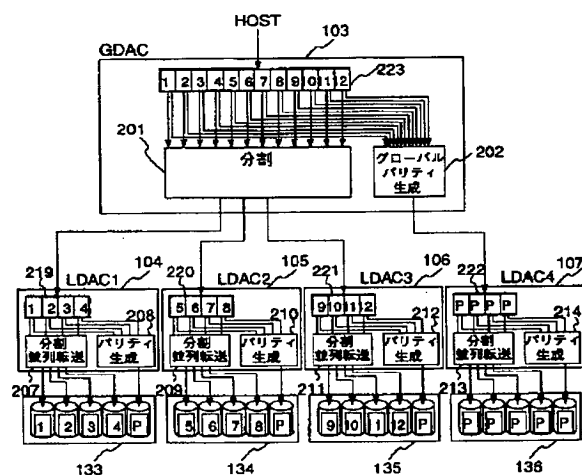
ラム、406…終了処理プログラム、1407…I/O要求受付処理プログラム、1408…データ回復処理プログラム、1409…I/O要求終了処理プログラム、1410…データ配置制御プログラム、1411…グローバルパリティ制御プログラム、1412…タグ制御プログラム、1405…パススイッチ。

【図2】



【図5】

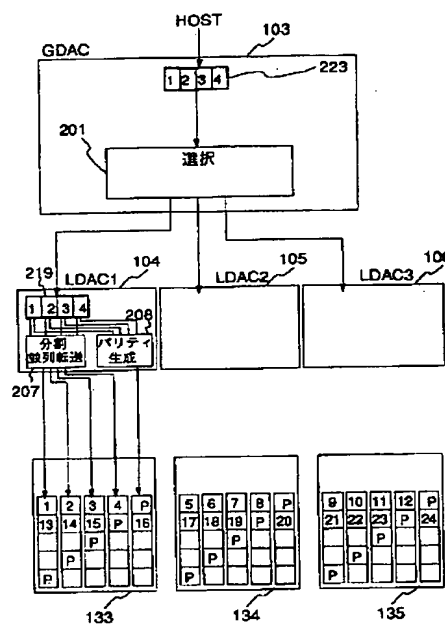
【図4】



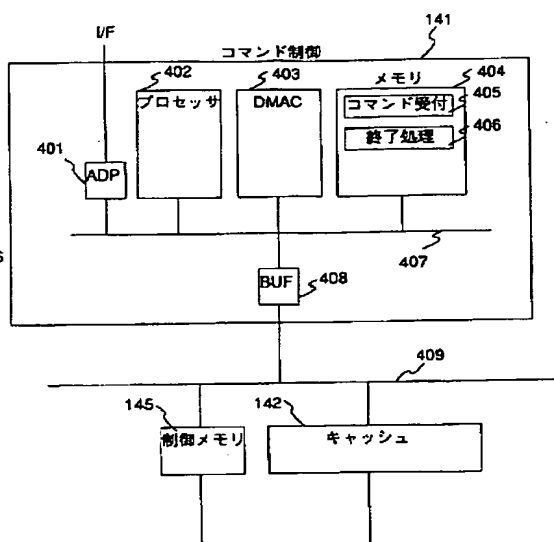
(21)

特開平7-200187

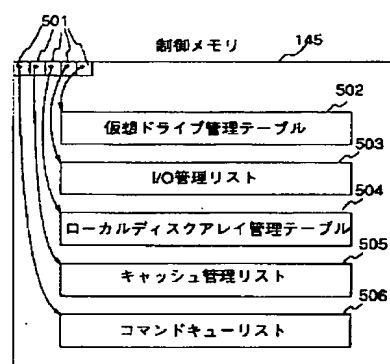
【図3】



【図6】



【図7】

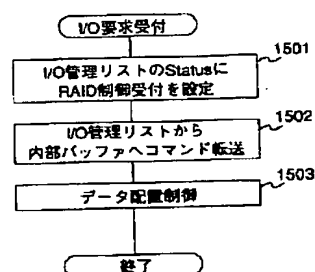


【図8】

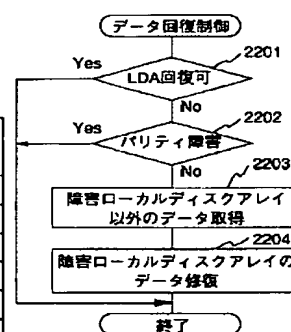
ローカルディスクアレイ管理テーブル

LDAC	PART	PASS	DRV	RAID	PRTY	ST-SCAP.	STATUS
LDAC1	1	1	5	5	1	1	Normal
	2	1	5	3	1	4	Normal
	3	1	5	0	1	1	Normal
LDAC2	1	2	5	5	1	1	Normal
	2	2	5	3	1	4	Normal
	3	2	5	0	1	1	Normal
LDAC3	1	3	5	5	1	1	Normal
	2	3	5	3	1	4	Normal
	3	3	5	0	1	1	Normal
LDAC4	1	4	5	5	1	1	Normal
	2	4	5	3	1	4	Normal
	3	4	5	0	1	1	Normal

【図17】



【図24】



(22)

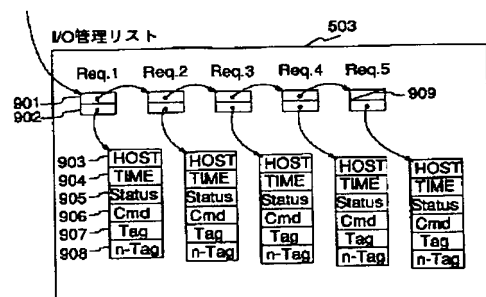
特開平7-200187

【図9】

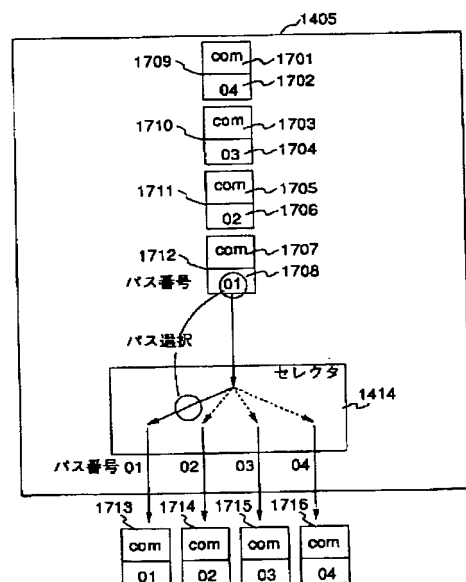
仮想ドライブ管理テーブル

ドライブ ID	ローカルディスクアレイ領域				Cap.	グローバル RAID	PRTY	ST-S	STATUS
PART	PART	PART	PART	PART					
VOL1	1	1	1	1	4	0	1	3	Normal
VOL2	2	2	2	2	4	3	1	16	Normal
VOL3	3	3	3	3	8	5	1	4	Normal

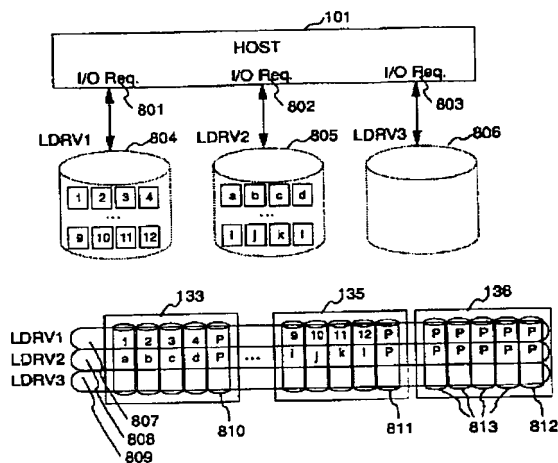
【図11】



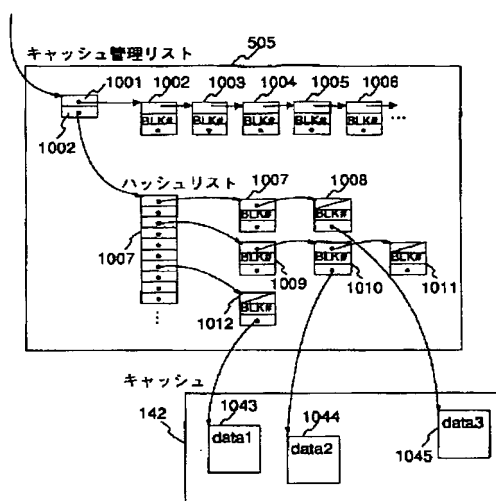
【図19】



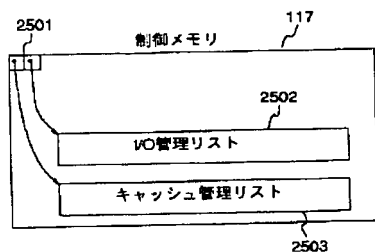
【図10】



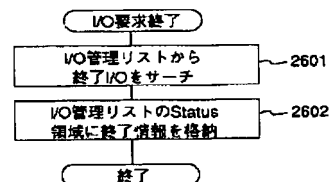
【図12】



【図27】



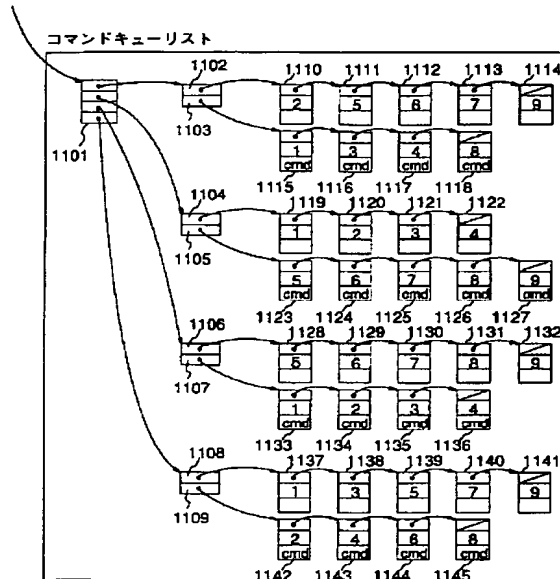
【図28】



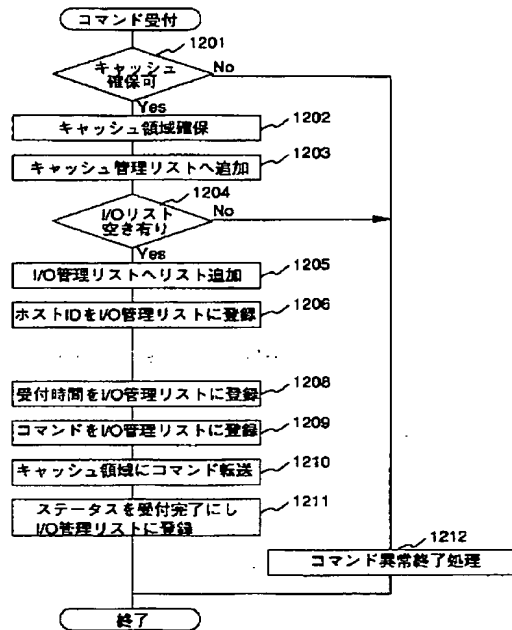
(23)

特開平7-200187

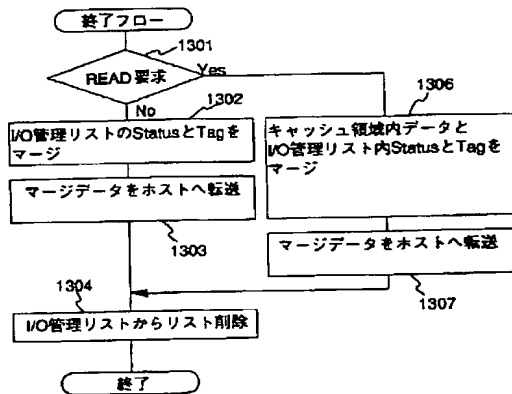
【図13】



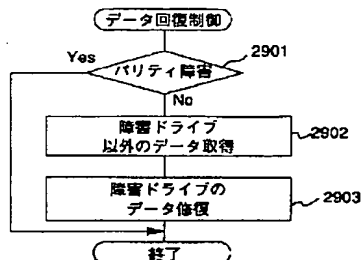
【図14】



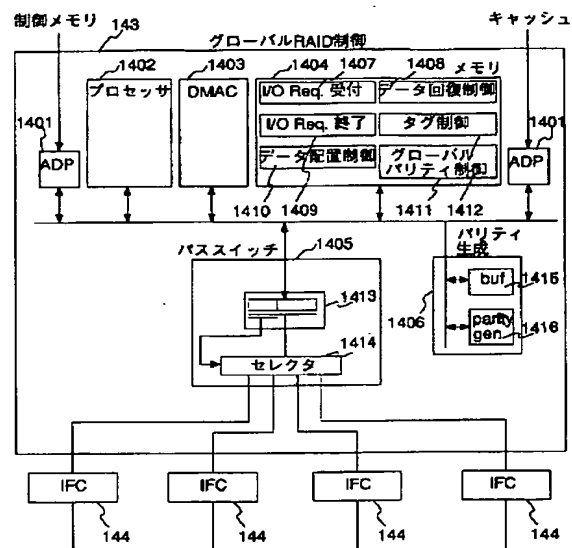
【図15】



【図31】



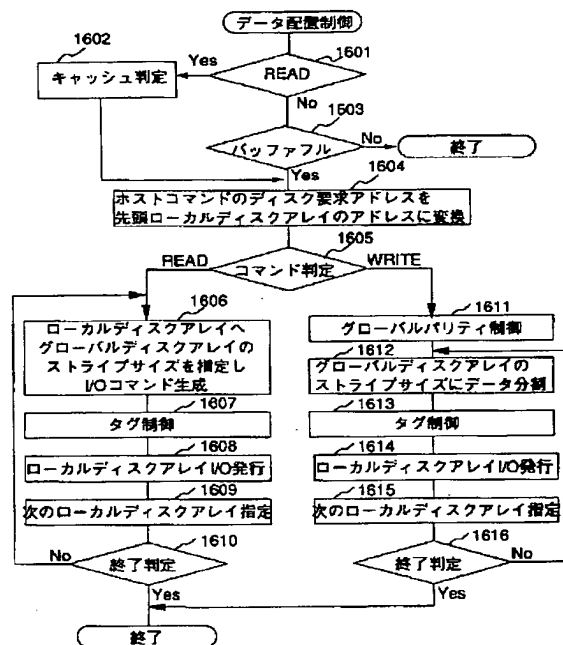
【図16】



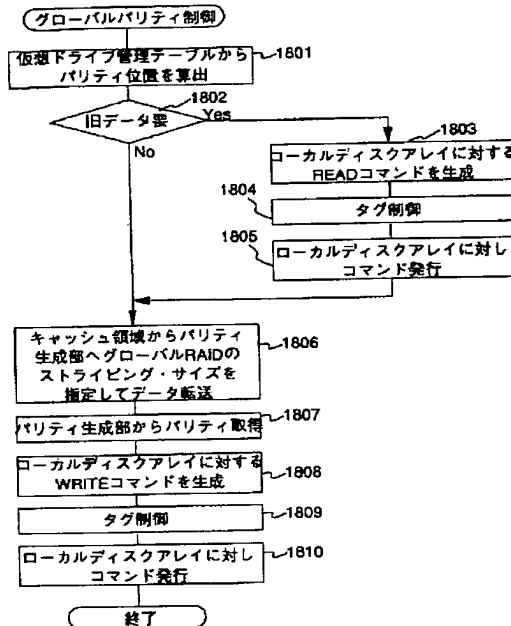
(24)

特開平7-200187

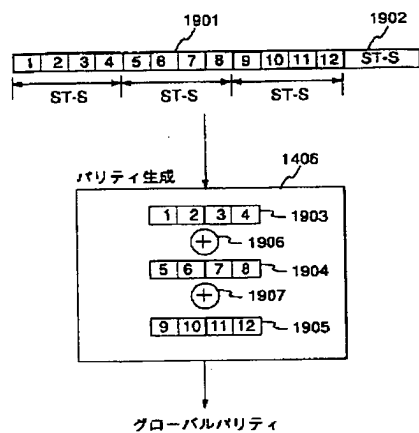
【図18】



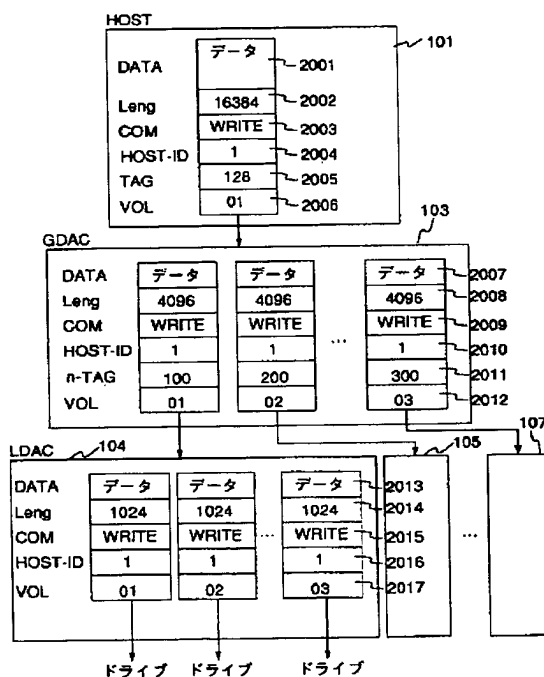
【図20】



【図21】



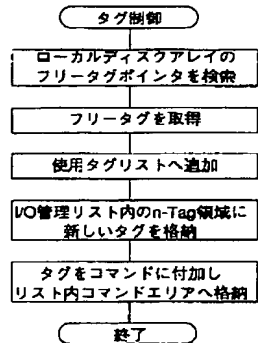
【図22】



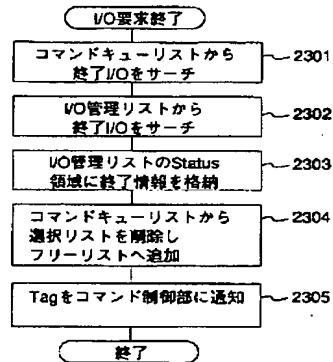
(25)

特開平7-200187

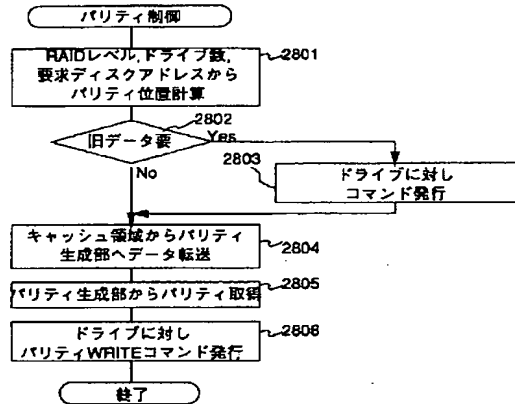
【図23】



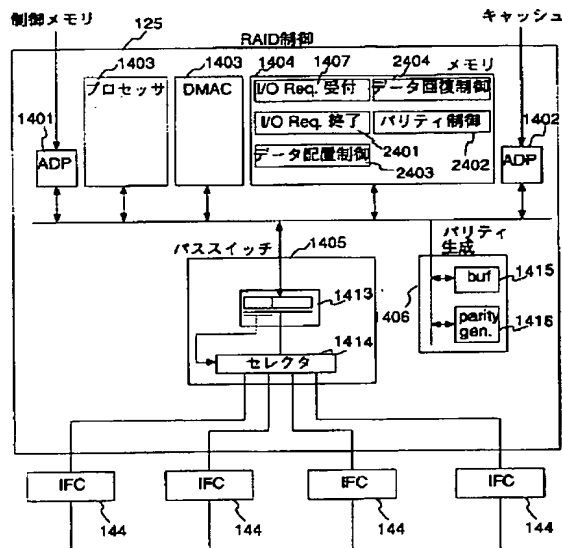
【図25】



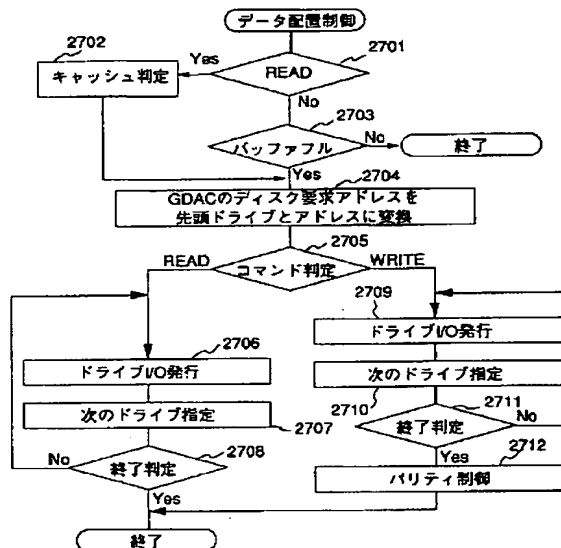
【図30】



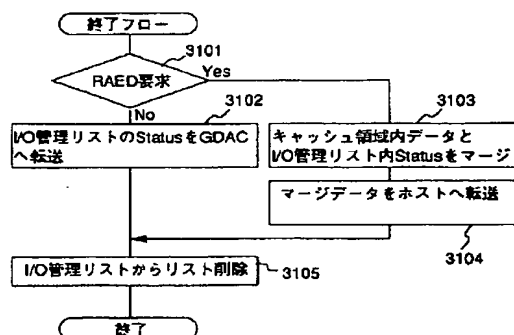
【図26】



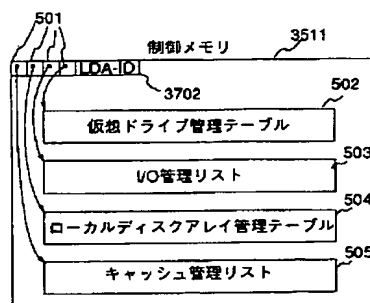
【図29】



【図33】



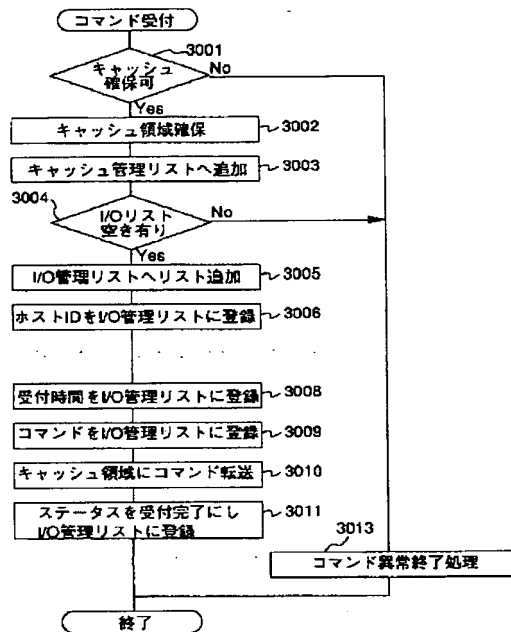
【図39】



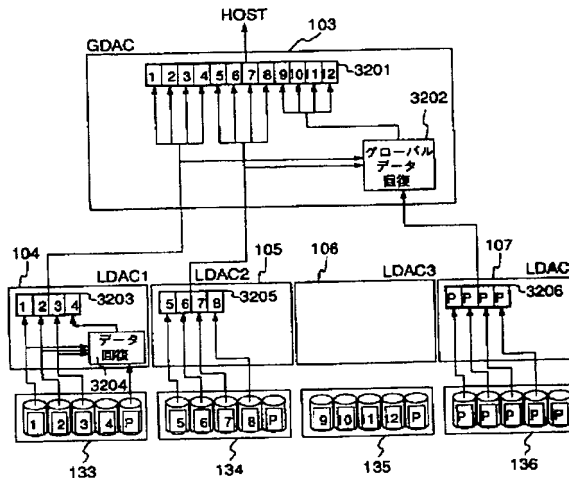
(26)

特開平 7-200187

【図 3 2】

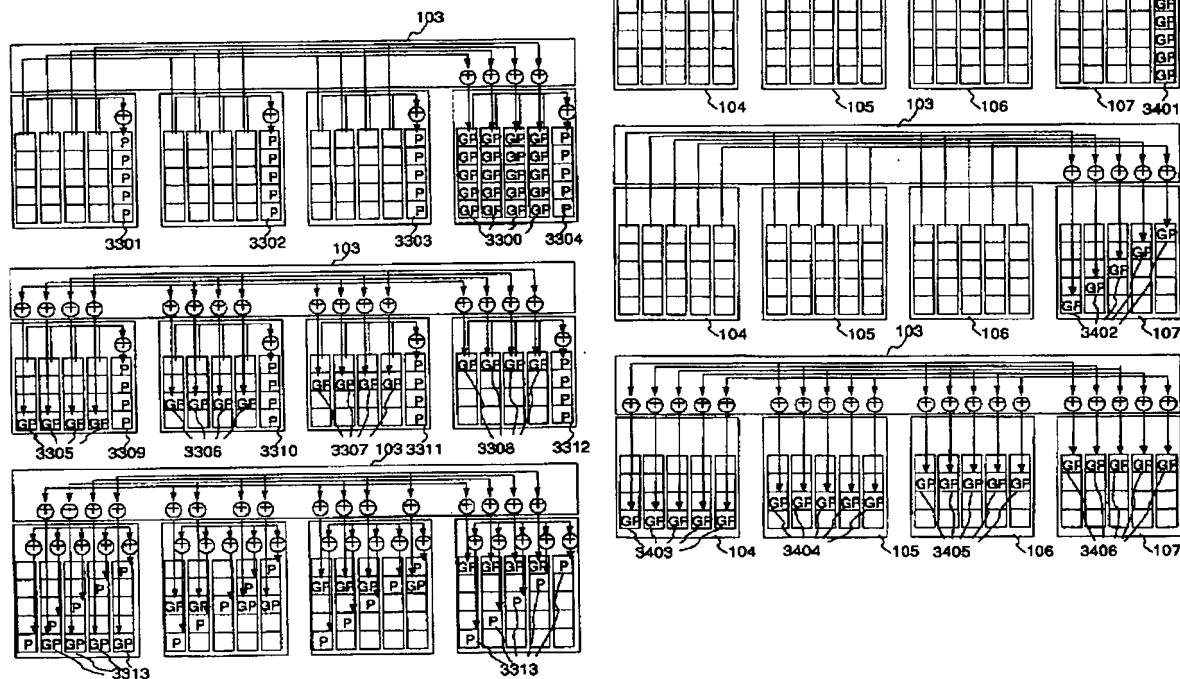


【図 3 4】



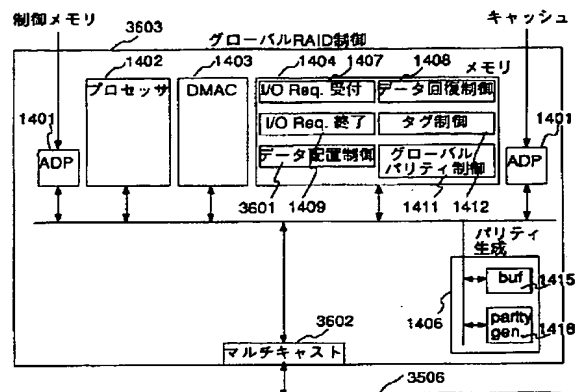
【図 3 6】

【図 3 5】

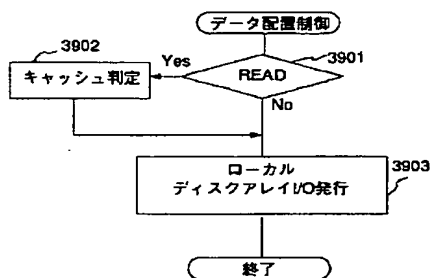


特開平7-200187

【图 3 8】



【图 4 1】



(28)

特開平7-200187

【図42】

